Hypothesis Testing: Concepts and Simple Examples

In this course, the statistical inference problems concerned with are inference problems regarding a parameter. A parameter can be estimated from sample data either by a single number (a point estimate) or an entire interval of plausible values (a confidence interval). Frequently, however, the objective of an investigation is not to estimate a parameter but to decide which of two contradictory claims about the parameter is correct. Methods for accomplishing this comprise the part of statistical inference is called *Hypothesis testing*.

1 Basic Concepts in Hypothesis Testing

Recall in our statistical inference problems, we are interested in the parameter θ of the probability distribution but the value of θ is unknown, we only know that the value of θ must lie in a certain parameter space Θ. We assume that Θ can be partitioned into two disjoint subsets Θ_0 and Θ_1, such that $\Theta_0 \cap \Theta_1 = \emptyset$ and $\Theta_0 \cup \Theta_1 = \Theta$. Now, we must decide whether the unknown value of θ lies in Θ_0 or in Θ_1.

Let H_0 denote the hypothesis that $\theta \in \Theta_0$ and let H_a denote the hypothesis that $\theta \in \Theta_1$. Since the subsets Θ_0 and Θ_1 are disjoint and their union is the whole parameter space, exactly one of the hypotheses H_0 and H_a must be true. We must decide whether to accept the hypothesis H_0 or the hypothesis H_a. A problem of this type, in which there are only two possible decisions, is called a problem of *hypothesis testing*.

In applications, we will make our decision based on some observations which are sampled from the probability distribution, and the observed values will provide us with information about the value of θ. A procedure to decide whether to accept the hypothesis H_0 or to accept the hypothesis H_a is called a *test procedure*.

In our discussion so far, we have treated the hypotheses H_0 and H_a on an equal base. In most problems, however, the two hypotheses are treated quite differently. In literature and in statistics community, the hypothesis H_0 is called *null hypothesis* and H_a is called *alternative hypothesis*.

Scientific research often involves trying to decide whether a current theory should be replaced by a more plausible and satisfactory explanation of the phenomenon under investigation. A conservative approach is to identify the current theory with H_0 and the researcher's alternative explanation with H_a. Rejection of the current theory will then occur only when evidence is much more consistent with the new theory. In many situations, H_a is referred as the "researcher's hypothesis" since it is the claim that the researcher would really like to validate. The word "null" means "of no value, effect or consequence" which suggests that H_0 should be identified with the hypothesis of no change (from current opinion), no difference, no improvement, and so on. Suppose, for example, that 10% of all circuit boards produced by a certain manufacturer during a recent period were defective. An engineer has suggested a change in the production process in the belief that it will result in a reduced defective rate. Let p denote the true proportion of defective boards resulting from the changed process. Then the research hypothesis, on which the burden of proof is placed, is the assertion that $p < 0.1$, thus the alternative hypothesis is $H_a : p < 0.1$. we usually don't say a theory is true!

Suppose that $X_1, \cdots, X_n$ form a random sample from a distribution for which the pdf or pf is $f(x|\theta)$, where the value of the parameter θ must lie in the parameter space Θ. Further assume that Θ_0 and Θ_1 are disjoint sets with $\Theta_0 \cup \Theta_1 = \Theta$. We want to test the following hypotheses:

$$H_0 : \theta \in \Theta_0 \quad \text{vs.} \quad H_a : \theta \in \Theta_1$$

The set Θ_i ($i = 0$ or 1) may contain just a single value of θ. If this is the case, then the corresponding hypothesis is said to be a *simple hypothesis*. On the other hand, if the set Θ_i contains more than one value of θ, then it is said that the corresponding hypothesis is a *composite hypothesis*. Under a simple hypothesis, the distribution of the observation is completely specified. Under a composite hypothesis, it is specified only that the distribution of the observations belongs to a certain class. For example, a simple null hypothesis H_0 must have the form

$$H_0 : \theta = \theta_0.$$

When θ is one-dimensional, there are two popular forms of composite hypotheses. *One-sided* null hypotheses are of the form $H_0 : \theta \le \theta_0$ or $H_0 : \theta \ge \theta_0$ with the corresponding one-sided alternative hypotheses being $H_a : \theta > \theta_0$ or $H_a : \theta < \theta_0$. When the null hypothesis is simple, like $H_0 : \theta = \theta_0$, the alternative hypothesis is usually *two-sided*, $H_a : \theta \ne \theta_0$.

In hypothesis testing problems, it is unavoidable to make mistakes, we might mistakenly reject or accept the null hypothesis. We should consider what kinds of errors we might make. For each value of $\theta \in \Theta_0$, the decision to reject H_0 is an incorrect decision. It has become traditional to call an erroneous decision to reject a true null hypothesis a *type I error*, or an error of the first kind. An erroneous decision to accept a false null hypothesis is called a *type II error*, or an error of the second kind. Of course, either $\theta \in \Theta_0$ or $\theta \in \Theta_1$, but not both. Hence, only one type of error is possible, but we never know which it is.

The probability of making a type I error is called the *significant level* of the test and is usually denoted as α; i.e.

$$\alpha = P(\text{reject } H_0 | H_0 \text{ true})$$

The probability of making a type II error is usually denoted as β, i.e.

$$\beta = P(\text{Accept } H_0 | H_a \text{ true}).$$

Ideally, we would like our decision makes mistake as small as possible, which means that we want α to be small and β to be small. However, in general, these two goals work against each other. That is, if we choose a decision rule to make α small, we will usually find β big. For example, the test procedure which always accept H_0, regardless of what data are observed, will have $\alpha = 0$. However, for this procedure, the type II error $\beta = 1$. Similarly, the test procedure which always reject H_0 will have $\beta = 0$, but always have $\alpha = 1$. Hence, there is a need to strike an appropriate balance between the two goals of small α and β.

In many practical applications type I errors are more delicate than type II errors (we should take care to reject a well-established theory). Therefore, the most popular method for striking a balance between the two goals is to control the type I error by choosing a number α_0 between 0 and 1, and require that $\alpha < \alpha_0$. This test is called a *level* α_0 *test*, and α_0 is called the *level of significance* for the test. While 1% or 5% might be an acceptable level of significance for one application, a different application can require a very different level.

Consider now a problem in which hypotheses having the following form are to be tested:

$$H_0 : \theta \in \Theta_0 \quad \text{vs.} \quad H_a : \theta \in \Theta_1$$

As we mentioned before, before we decide which hypothesis to accept, we can observe a random sample $X_1, \cdots, X_n$ drawn from a distribution that involves the unknown parameter θ. We let S denote the sample space of the n-dimensional random vector $\mathbf{X} = (X_1, \cdots, X_n)$. In other words, S is the set of all possible outcomes of the random sample.

In the problem of this type, we specify a test procedure by partitioning the sample space S into two subsets. One subset contains the values of $\mathbf{X}$ for which the test procedure will accept H_0, and the other subset contains the values of $\mathbf{X}$ for which the test procedure will reject H_0 and therefore accept H_a. The subset for which H_0 will be rejected is called the *critical region* of the test. Therefore, the test procedure is determined by specifying the critical region of the test, and the complement of the critical region must then contain all the outcomes for which H_0 will be accepted.

We usually find a function of the sample and make our decision based on the value of the function. Therefore the function should be computable, so it does not contain any unknown parameter. In another word, the function based on which we make our decision to reject or accept the null hypothesis is a statistic, and this statistic is called a *testing statistic*. Let us denote this test statistic as $T = r(\mathbf{X})$. Under the null hypothesis, the probability distribution of the test statistic is called the *null distribution*.

Typically, a test using test statistic T will reject the null hypothesis if T falls in some fixed interval or falls outside of some fixed interval. The set of values of the test statistic that leads to rejection of the null hypothesis is called the *rejection region*, and the set of values that leads to acceptance of the null hypothesis is called the *acceptance region*.

Let us study a simple example which demonstrates the above mentioned ideas.

Example 1: A politician claimed that he will gain 50% of votes in a city election and thereby he will win. However, after studied the politician's policies, an expert thinks he will lose the election. To test their conjecture, the expert randomly select 15 voters and found Y of them would vote for the candidate.

Let us denote the supporting rate for the candidate is p. The researcher's hypothesis is the expert's conjecture, $p < 0.5$, and the null hypothesis is the politician's claim, $p \geq 0.5$. Therefore, we have the following hypotheses:

$$H_0 : p \geq 0.5 \quad \text{vs.} \quad H_a : p < 0.5.$$

We are given that Y out of 15 voters support the candidate. If the value of Y is very small, say 0, what would we conclude about the candidate's claim? If the candidate was correct, i.e., there are at least 50% of voters support him, it is not impossible to observe $Y = 0$ people favoring the candidate out of 15 voters, but the possibility should be very small. It is much more likely that we would observe $Y = 0$ if the alternative hypothesis was correct. Thus, if we observe $Y = 0$, intuitively, we would reject the null hypothesis ($p \geq 0.5$) and in favor of the alternative hypothesis ($p < 0.5$). If we observe other small values of Y, the similar reasoning would lead us to the same conclusion.

In this example, we make our decision based on the value of Y, therefore, Y is the test statistic. If the null hypothesis was correct, i.e. supporting rate $p \geq 0.5$, Y is distributed as a binomial distribution, i.e.

$$Y \sim \text{Bin}(15, p).$$

We should select a value for p to fully determine the null distribution. A general rule is to select p in H_0 and it is closest to H_a, here obviously we select $p = 0.5$. Thus, $\text{Bin}(15, p)$ is the null distribution.

From our discussion above, we would reject the null hypothesis if the observed value Y is too small, therefore the rejection region is of the form $\{Y \leq c\}$, where c is a constant.

Let us calculate the probability of type I error if we select the rejection region as $\{Y \leq 2\}$. By definition,

$$\alpha = P(Y \leq 2|H_0) = \sum_{y=0}^{2} \binom{15}{y} (0.5)^y (0.5)^{15-y} = \binom{15}{0} (0.5)^{15} + \binom{15}{1} (0.5)^{15} + \binom{15}{2} (0.5)^{15} = 0.004.$$

Suppose that the candidate will receive 30% of the votes ($p = 0.3$), calculate the probability β that the sample will erroneously lead us to conclude that H_0 is true. This probability is the probability of type II error. By definition,

$$\beta = P(\text{accept } H_0 | H_a \text{ true}) = P(Y > 2|p = 0.3) = \sum_{y=3}^{15} \binom{15}{y} (0.3)^y (0.7)^{15-y} = 0.873,$$

because when H_a is true, $Y \sim \text{Bin}(15, 0.3)$.

Now, assume we adjusted the reject region to be $\{Y \leq 5\}$, and still assume the true value of p is 0.3, let us calculate the probability of type I and type II errors α and β. In this case

$$\alpha = P(Y \leq 5|H_0) = \sum_{y=0}^{5} \binom{15}{y} (0.5)^y (0.5)^{15-y} = \sum_{y=0}^{5} \binom{15}{y} (0.5)^{15} = 0.151.$$

and

$$\beta = P(\text{accept } H_0|H_a \text{ true}) = P(Y > 5|p = 0.3) = \sum_{y=6}^{15} \binom{15}{y} (0.3)^y (0.7)^{15-y} = 0.278.$$

This example shows that the test using rejection region $\{Y \leq 2\}$ can guarantee a low risk of making a type I error ($\alpha = 0.004$), but it does not offer adequate protection against a type II error ($\beta = 0.873$). We can decrease the probability of type II error to 0.278 by changing the rejection region to $\{Y \leq 5\}$, however, by doing this we increased the probability of type I error to 0.151. Thus, this example demonstrates the inverse relation between α and β.

2 Some Commonly used Tests

This section gives some commonly used, canonical hypothesis testing procedures.

Example 2: Testing Hypotheses about the mean of a normal distribution with known variance. Suppose that $\mathbf{X} = (X_1, \cdots, X_n)$ is a random sample from a normal distribution with unknown mean μ and known variance σ^2. We wish to test the hypotheses

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_a : \mu \neq \mu_0$$

Solution: If the null hypothesis was correct, we would expect the sample mean value $\bar{X}$ to be close to the population mean value μ_0. Under the null hypothesis, we know that $\bar{X} \sim N(\mu_0, \sigma^2/n)$, that is

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1).$$

If the null hypothesis was correct, then the statistic Z should be close to 0. We make our decision based on the value of Z, therefore Z is the test statistic, and the distribution of Z under the assumption of H_0 is the null distribution, in this example, the null distribution is $N(0, 1)$.

Now suppose we want to test the hypotheses on the significant level α, that is under the null hypothesis $\mu = \mu_0$, the probability of rejecting H_0 is α. As we analyzed before, we make our decision based on the deviation of Z from 0, that is if the value of $|Z| > c$, we will reject the null hypothesis, where c is a constant. Therefore the rejection probability is $P(|Z| > c) = \alpha$ where $Z \sim N(0, 1)$. It is easy to calculate that the value of c is $100 \times (1-\alpha/2)$ percentile of standard normal distribution, $Z(1 - \alpha/2)$. Therefore, the final decision rule is: if $|Z| > Z(1-\alpha/2)$, we would reject the null hypothesis at the significant level α; otherwise, we fail to reject the null hypothesis.

In the above, the alternative hypothesis is two-sided: $\mu \neq \mu_0$. If the alternative is one-sided, for example,

$$H_0 : \mu \leq \mu_0 \quad \text{vs.} \quad H_a : \mu > \mu_0,$$

then we would reject the null hypothesis if the test statistic Z is greater than a predetermined positive number c. If the significant level is α, we can calculate the value of c by $P(Z > c) = \alpha$ where $Z \sim N(0, 1)$. Then $c = Z(1-\alpha)$. Therefore if the alternative hypothesis is $H_a : \mu > \mu_0$, we can reject the null hypothesis if the test statistic $Z > Z(1 - \alpha)$.

The other case of one-sided hypothesis is

$$H_0 : \mu \geq \mu_0 \quad \text{vs.} \quad H_a : \mu < \mu_0.$$

Similarly we would reject the null hypothesis if the test statistic Z is less than a predetermined negative number c. If the significant level is α, we can calculate the value of c by $P(Z < c) = \alpha$ where $Z \sim N(0, 1)$. Then $c = Z(\alpha)$. Therefore if the alternative hypothesis is $H_a : \mu \leq \mu_0$, we can reject the null hypothesis if the test statistic $Z < Z(\alpha)$.

Example 3: Testing Hypotheses about the mean of a normal distribution with unknown variance. Suppose that $\mathbf{X} = (X_1, \cdots, X_n)$ is a random sample from a normal distribution with unknown mean μ and unknown variance σ^2. We wish to test the hypotheses

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_a : \mu \neq \mu_0$$

Solution: Based on the same reasoning as in Example 2, if the null hypothesis was correct, we would expect the sample mean value $\bar{X}$ to be close to the population mean value μ_0. However, in this case, we cannot use

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

as the test statistic because σ is unknown. However, if we use

$$S = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2}$$

to replace σ, we have

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

is a statistic, and the distribution of T is t distribution with $n-1$ degrees of freedom, i.e., under the null hypothesis

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}.$$

Therefore, in this example, T is the test statistic, and null distribution is t_{n-1}.

Now suppose we want to test the hypotheses on the significant level α, that is under the null hypothesis $\mu = \mu_0$, the probability of rejecting H_0 is α. As we analyzed before, we make our decision based on the deviation of T from 0, that is if the value of $|T| > c$, we will reject the null hypothesis, where c is a constant. Therefore the rejection probability is $P(|T| > c) = \alpha$ where $T \sim t_{n-1}$. It is easy to calculate that the value of c is $100 \times (1 - \alpha/2)$ percentile of t_{n-1} distribution, $t_{n-1}(1 - \alpha/2)$. Therefore, the final decision rule is: if $|T| > t_{n-1}(1 - \alpha/2)$, we would reject the null hypothesis at the significant level α; otherwise, we fail to reject the null hypothesis.

Now let us consider the one-sided test

$$H_0 : \mu \leq \mu_0 \quad \text{vs.} \quad H_a : \mu > \mu_0.$$

In this case, we would reject the null hypothesis if the test statistic T is greater than a predetermined positive number c. If the significant level is α, we can calculate the value of c by $P(T > c) = \alpha$ where $T \sim t_{n-1}$. Then $c = t_{n-1}(1 - \alpha)$. Therefore if the alternative hypothesis is $H_a : \mu > \mu_0$, we can reject the null hypothesis if the test statistic $T > t_{n-1}(1 - \alpha)$.

The other case of one-sided hypothesis is

$$H_0 : \mu \geq \mu_0 \quad \text{vs.} \quad H_a : \mu < \mu_0.$$

Similarly we would reject the null hypothesis if the test statistic T is less than a predetermined negative number c. If the significant level is α, we can calculate the value of c by $P(T < c) = \alpha$ where $T \sim t_{n-1}$. Then $c = t_{n-1}(\alpha)$. Therefore if the alternative hypothesis is $H_a : \mu \leq \mu_0$, we can reject the null hypothesis if the test statistic $T < t_{n-1}(\alpha)$.

Example 4: Suppose $X_1, \cdots, X_n$ from a normal distribution $N(\mu, \sigma^2)$ where μ is known and σ is unknown. We wish to test the hypotheses

$$H_0 : \sigma = \sigma_0 \quad \text{vs.} \quad H_a : \sigma \neq \sigma_0$$

Solution: We use $\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$ to estimate σ^2. If the null hypothesis was correct, the estimated value should be close to the value of σ_0^2. Enlightened by this intuitive idea, we would reject the null hypothesis if either $\widehat{\sigma^2}/\sigma_0^2$ is too big or too small. Recall that the statistic

$$V = \frac{n\widehat{\sigma^2}}{\sigma_0^2}$$

follows chi-square distribution with n degrees of freedom. We can use V as the test statistic, and the null distribution under the null hypothesis is χ^2_n.

Now, Suppose we want to test the hypotheses at the significant level α. From our intuition, the probability of rejecting the null hypothesis is

$$P(V < a \text{ or } V > b) = \alpha$$

Where a and b are two constants. Let $\chi^2_n(\alpha/2)$ and $\chi^2_n(1-\alpha/2)$ be the $(\alpha/2) \times 100$-th and $(1-\alpha/2) \times 100$-th percentiles, respectively. We can let $a = \chi^2_n(\alpha/2)$ and $b = \chi^2_n(1-\alpha/2)$, respectively. Therefore, the decision rule would be: if $V < \chi^2_n(\alpha/2)$ or $V > \chi^2_n(1-\alpha/2)$, we would reject the null hypothesis; otherwise we accept the null hypothesis.

Now let us consider the one-sided test

$$H_0 : \sigma \leq \sigma_0 \quad \text{vs.} \quad H_a : \sigma > \sigma_0.$$

In this case, we would reject the null hypothesis if the test statistic V is greater than a predetermined positive number b. If the significant level is α, we can calculate the value of b by $P(V > b) = \alpha$ where $V \sim \chi^2_n$. Then $b = \chi^2_n(1-\alpha)$. Therefore if the alternative hypothesis is $H_a : \sigma > \sigma_0$, we can reject the null hypothesis if the test statistic $T > \chi^2_n(1-\alpha)$.

The other case of one-sided hypothesis is

$$H_0 : \sigma > \sigma_0 \quad \text{vs.} \quad H_a : \sigma \leq \sigma_0.$$

Similarly we would reject the null hypothesis if the test statistic V is less than a predetermined number a. If the significant level is α, we can calculate the value of a by $P(V < a) = \alpha$ where $V \sim \chi^2_n$. Then $a = \chi^2_n(\alpha)$. Therefore if the alternative hypothesis is $H_a : \sigma \leq \sigma_0$, we can reject the null hypothesis if the test statistic $V < \chi^2_n(\alpha)$.

Summarizing the examples above, we can see that the general procedure for a hypothesis testing problem is like the following steps:

1. Find an estimator for the parameter of interest.
2. Find a connection between the estimator and the given value of the parameter in H_0. This step gives us an intuitive idea for the testing procedure, but not necessarily gives us the test statistic.
3. Usually, last step gives us several candidates for the test statistic, choose the one which yields a standard distribution (modify the test statistic if necessary).
4. Determine the rejection region based on the analysis, and express the type I error rate based on the null distribution.
5. Solve the obtained equation to find the decision rule.

Let us do the following example following the above steps.

Example 5: Suppose $X_1, \cdots, X_n$ from a normal distribution $N(\mu, \sigma^2)$ where both μ and σ are unknown. We wish to test the hypotheses

$$H_0 : \sigma = \sigma_0 \quad \text{vs.} \quad H_a : \sigma \neq \sigma_0$$

Step 1: In the case of both μ and σ are unknown, we would estimate σ^2 by $\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$.

Step 2: If the null hypothesis was correct, we would expect that the estimated value $\widehat{\sigma^2}$ would be close to σ_0^2, and this suggest us to compare $\widehat{\sigma^2}$ and σ_0^2. Therefore, the test statistic could involve quantities like $\widehat{\sigma^2} - \sigma_0^2$ or $\widehat{\sigma^2}/\sigma_0^2$, and so on.

Step 3: Recall that

$$V' = \frac{n\widehat{\sigma^2}}{\sigma_0^2} \sim \chi^2_{n-1}$$

Therefore, in this example, V' is the test statistic, and null distribution is χ^2_{n-1}. As you can see that we choose $\widehat{\sigma^2}/\sigma_0^2$ from the two candidates in Step 2, and modified it so that it has a standard distribution.

Step 4: Now suppose we want to test the hypotheses on the significant level α, that is under the null hypothesis $\sigma = \sigma_0$, the probability of rejecting H_0 is α. As we analyzed before, we make our decision based on the value of V', that is if the value of $V' < a$ or $V' > b$, we will reject the null hypothesis, where a and b are constants. Therefore the rejection probability is $P(V' < a \text{ or } V' > b) = \alpha$ where $V' \sim \chi^2_{n-1}$.

Step 5: Solving the equation in step 4, we can let $a = \chi^2_{n-1}(\alpha/2)$ and $b = \chi^2_{n-1}(1-\alpha/2)$, respectively. Therefore, the decision rule would be: if $V' < \chi^2_{n-1}(\alpha/2)$ or $V' > \chi^2_{n-1}(1-\alpha/2)$, we would reject the null hypothesis; otherwise we accept the null hypothesis.

Now let us consider the one-sided test

$$H_0 : \sigma \leq \sigma_0 \quad \text{vs.} \quad H_a : \sigma > \sigma_0.$$

In this case, we would reject the null hypothesis if the test statistic V' is greater than a predetermined positive number b. If the significant level is α, we can calculate the value of b by $P(V' > b) = \alpha$ where $V \sim \chi^2_{n-1}$. Then $b = \chi^2_{n-1}(1-\alpha)$. Therefore if the alternative hypothesis is $H_a : \sigma > \sigma_0$, we can reject the null hypothesis if the test statistic $V' > \chi^2_{n-1}(1-\alpha)$.

The other case of one-sided hypothesis is

$$H_0 : \sigma > \sigma_0 \quad \text{vs.} \quad H_a : \sigma \leq \sigma_0.$$

Similarly we would reject the null hypothesis if the test statistic V' is less than a predetermined number a. If the significant level is α, we can calculate the value of a by $P(V' < a) = \alpha$

where $V' \sim \chi^2_{n-1}$. Then $a = \chi^2_{n-1}(\alpha)$. Therefore if the alternative hypothesis is $H_a : \sigma \leq \sigma_0$, we can reject the null hypothesis if the test statistic $V' < \chi^2_n(\alpha)$.

3 Attained Significance Level, or p-Value

As we discussed before, the type I error probability, α, is usually called the significance level, or the level of the test. However, in applications, the actually used value of α is somewhat arbitrary. Thus, it is possible that we will have different conclusions for $\alpha = 0.05$ and $\alpha = 0.01$. Although, in applications, people usually use $\alpha = 0.05$ or $\alpha = 0.01$ frequently, this choice is for the sake of convenience rather than as a result of rigorous consideration. Thus, we need a more informative quantity than the statement like "we reject the null hypothesis at the level 0.05".

Let W be a test statistic, the *p-value*, or *attained significance level*, is the smallest level of significance α for which the observed data indicate that the null hypothesis should be rejected.

The smaller the p-value is, the more convincing is the evidence that the null hypothesis should be rejected. If an experimenter has a value of α in mind, the p-value can be used to implement an α-level test. Since the p-value is the smallest value of α for which the null hypothesis can be rejected, thus if the desired value of α is greater than or equal to the p-value, the null hypothesis should be rejected for that value of α. Indeed, the null hypothesis should be rejected for any value of α down to and including the p-value. Otherwise, if α is less than the p-value, the null hypothesis cannot be rejected. In a sense, the p-value allows the reader to evaluate the extent to which the observed data disagree with the null hypothesis. Thus, p-value is informative.

If we were to reject the null hypothesis H_0 in favor of the alternative hypothesis H_a for small values of a test statistic W, that is, the rejection region is $\{W \leq c\}$, then the p-value associated with an observed value w_0 of W could be calculated as

$$p\text{-value} = P(W \leq w_0 | H_0 \text{ true}).$$

Similarly, if the rejection region is of the form $\{W \geq c\}$, then the p-value associated with an observed value w_0 of W could be calculated as

$$p\text{-value} = P(W \geq w_0 | H_0 \text{ true}).$$

For two-sided test, suppose the rejection region is of the form $\{|W| \geq c\}$, then the p-value associated with an observed value w_0 of W could be calculated as

$$p\text{-value} = P(|W| \geq w_0 | H_0 \text{ true}) = P(W \geq w_0 | H_0 \text{ true}) + P(W \leq -w_0 | H_0 \text{ true}).$$

We can see that the p-value actually is the tail probability of the null distribution: if the test is a one-sided test, then the p-value is one-tail probability; if the test is a two-sided test, then the p-value is two-tail probability.

Example 6: Testing Hypotheses about the mean of a normal distribution with unknown variance. Suppose that $\mathbf{X} = (X_1, \cdots, X_n)$ is a random sample from a normal distribution with unknown mean μ and unknown variance σ^2. Please find the p-value for the hypotheses testing:

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_a : \mu \neq \mu_0$$

Solution: From Example 3, we know the rejection region for this test is of the form $\{|T| > c\}$, where

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1},$$

and

$$S = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2}.$$

Now suppose from the observation $x_1, \cdots, x_n$, we calculate the value of the test statistic as t^*, then the p-value can be calculated as

$$p\text{-value} = P(|T| > t^*),$$

where $T \sim t_{n-1}$. Let the desired significance level be α, then if p-value $< \alpha$, we reject the null hypothesis; otherwise we fail to reject the null hypothesis.

Example 7: Suppose $X_1, \cdots, X_n$ from a normal distribution $N(\mu, \sigma^2)$ where both μ and σ are unknown. Find the p-value for the hypotheses testing problem:

$$H_0 : \sigma = \sigma_0 \quad \text{vs.} \quad H_a : \sigma > \sigma_0$$

Solution: From Example 5, we will have

$$V = \frac{n\widehat{\sigma^2}}{\sigma_0^2} \sim \chi^2_{n-1},$$

where $\widehat{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$. The rejection region is of the form $V > c$.

Now suppose from the observation $x_1, \cdots, x_n$, we calculate the value of the test statistic as v^*, then the p-value can be calculated as

$$p\text{-value} = P(V > v^*),$$

where $V \sim \chi^2_{n-1}$. Let the desired significance level be α, then if p-value $< \alpha$, we reject the null hypothesis; otherwise we fail to reject the null hypothesis.

4 Calculating Type II Error Probability and Finding the Sample Size for a Normal Distribution with Known Variance

Calculating the type II error probability β could be very difficult for some statistical tests, but it is easy for the case when the variance is known in the normal distribution.

Suppose that $\mathbf{X} = (X_1, \cdots, X_n)$ is a random sample from a normal distribution with unknown mean μ and known variance σ^2. We wish to test the hypotheses

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_a : \mu \neq \mu_0.$$

at the significant level α.

From Example 2 in this section, we know that, under the null hypothesis

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1).$$

and if $|Z| > Z(1 - \alpha/2)$, we would reject the null hypothesis at the significant level α; otherwise, we fail to reject the null hypothesis. That is, we would reject the null hypothesis, if

$$\bar{X} > \mu_0 + z(1 - \alpha/2)\frac{\sigma}{\sqrt{n}} \quad \text{or} \quad \bar{X} < \mu_0 - z(1 - \alpha/2)\frac{\sigma}{\sqrt{n}}.$$

Suppose the alternative hypothesis is true, i.e. the true value of μ is some value $\mu_a \neq \mu_0$, let us calculate the probability of type II error for this test. Under the alternative hypothesis, we have

$$Z' = \frac{\bar{X} - \mu_a}{\sigma/\sqrt{n}} \sim N(0, 1).$$

By definition,

$$\begin{aligned}
\beta &= P(\text{Accept } H_0 | H_a \text{ true}) \\
&= P\left(\mu_0 - z(1 - \alpha/2)\frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq \mu_0 + z(1 - \alpha/2)\frac{\sigma}{\sqrt{n}} | \mu = \mu_a\right) \\
&= P\left(\frac{\mu_0 - z(1 - \alpha/2)\sigma/\sqrt{n} - \mu_a}{\sigma/\sqrt{n}} \leq \frac{\bar{X} - \mu_a}{\sigma/\sqrt{n}} \leq \frac{\mu_0 + z(1 - \alpha/2)\sigma/\sqrt{n} - \mu_a}{\sigma/\sqrt{n}}\right) \\
&= P\left(\frac{\mu_0 - \mu_a}{\sigma/\sqrt{n}} - z(1 - \alpha/2) \leq Z' \leq \frac{\mu_0 - \mu_a}{\sigma/\sqrt{n}} + z(1 - \alpha/2)\right) \\
&= \Phi\left(\frac{\mu_0 - \mu_a}{\sigma/\sqrt{n}} + z(1 - \alpha/2)\right) - \Phi\left(\frac{\mu_0 - \mu_a}{\sigma/\sqrt{n}} - z(1 - \alpha/2)\right)
\end{aligned}$$

where $\Phi(z)$ is the cumulative distribution function of the standard normal random variable.

The above example also suggests a procedure for us to determine the sample size for an experiment. Suppose for the sample above, we want to test hypotheses

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_a : \mu > \mu_0,$$

and we specify the value of desired α and β. β is evaluated when $\mu = \mu_a > \mu_0$, where μ_a is the true value of μ. Let us determine the sample size n.

Intuitively, if the sample mean $\bar{X}$ is big enough, then we should reject the null hypothesis, therefore the rejection region is of the form $\{\bar{X} > c\}$ where c is a constant. If the null hypothesis was true, we have

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1).$$

Under the alternative hypothesis, we have

$$Z' = \frac{\bar{X} - \mu_a}{\sigma/\sqrt{n}} \sim N(0, 1).$$

By the definitions of α and β, we have

$$\begin{aligned} \alpha &= P(\text{Reject } H_0 | H_0 \text{ true}) = P(\bar{X} > c | \mu = \mu_0) \\ &= P\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > \frac{c - \mu_0}{\sigma/\sqrt{n}} | \mu = \mu_0\right) = P(Z > z(1 - \alpha)) \end{aligned}$$

and

$$\begin{aligned} \beta &= P(\text{Accept } H_0 | H_a \text{ true}) = P(\bar{X} \leq c | \mu = \mu_a) \\ &= P\left(\frac{\bar{X} - \mu_a}{\sigma/\sqrt{n}} \leq \frac{c - \mu_a}{\sigma/\sqrt{n}} | \mu = \mu_a\right) = P(Z' < z(\beta)) \end{aligned}$$

The above equations give us new equations

$$\frac{c - \mu_0}{\sigma/\sqrt{n}} = z(1 - \alpha) \quad \text{and} \quad \frac{c - \mu_a}{\sigma/\sqrt{n}} = z(\beta).$$

Solving for c gives us

$$c = \mu_0 + \frac{\sigma}{\sqrt{n}} z(1 - \alpha) = \mu_a + \frac{\sigma}{\sqrt{n}} z(\beta),$$

thus,

$$\sqrt{n} = \frac{\sigma(z(1 - \alpha) - z(\beta))}{\mu_a - \mu_0}.$$

Finally, the sample size for an upper-tail α-level test is

$$n = \frac{\sigma^2 (z(1 - \alpha) - z(\beta))^2}{(\mu_a - \mu_0)^2}.$$

5 Duality of Confidence Intervals and Hypothesis Tests

We have studied the confidence intervals for a parameter, and the hypothesis testing techniques regarding the unknown parameters. In the previous examples, we can see that for an unknown parameter, we used the same sampling distribution to get a confidence interval and do hypothesis testing. Actually, there is a duality between confidence intervals and hypothesis tests: a confidence interval can be obtained by "inverting" a hypothesis test, and vice versa. We will demonstrate this by several examples.

Example 8: Suppose $X_1, \cdots, X_n$ from a normal distribution $N(\mu, \sigma^2)$ where μ is unknown but σ is known. Let us consider the following hypothesis testing problem:

$$H_0 : \mu = \mu_0$$
$$H_a : \mu \neq \mu_0$$

For this hypothesis testing problem, we use the test statistic

$$\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$$

Let $Z(1-\alpha/2)$ be the $(1-\alpha/2)$ percentile, from the previous discussion, we know that we would accept H_0 if

$$\left|\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}\right| \leq Z(1-\alpha/2)$$

or

$$-Z(1-\alpha/2)\frac{\sigma}{\sqrt{n}} \leq \bar{X} - \mu_0 \leq Z(1-\alpha/2)\frac{\sigma}{\sqrt{n}}$$

or

$$\bar{X} - Z(1-\alpha/2)\frac{\sigma}{\sqrt{n}} \leq \mu_0 \leq \bar{X} + Z(1-\alpha/2)\frac{\sigma}{\sqrt{n}} \tag{1}$$

The hypothesis testing procedure says that if $\bar{X}$ and μ_0 satisfies the inequality (1), we should accept the null hypothesis.

We can recognize that the interval

$$\left[\bar{X} - Z(1-\alpha/2)\frac{\sigma}{\sqrt{n}}, \bar{X} + Z(1-\alpha/2)\frac{\sigma}{\sqrt{n}}\right] \tag{2}$$

is the $(1-\alpha)$ confidence interval for the parameter μ. The above hypothesis testing procedure says that we can get the $1-\alpha$ confidence interval from the result of level α significant hypothesis testing.

Conversely, from the $1-\alpha$ confidence interval (2), we can see that if the value of μ_0 is in this interval, then we accept H_0; otherwise, we reject H_0. And this is exactly the decision rule expressed by the hypothesis testing result.

Example 9: Suppose $X_1, \cdots, X_n$ from a normal distribution $N(\mu, \sigma^2)$ where μ is known and σ^2 is unknown. We want to do the following hypothesis testing:

$$H_0 : \sigma^2 = \sigma_0^2$$
$$H_a : \sigma^2 \neq \sigma_0^2$$

We use $\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2$ to estimate σ^2. For this hypothesis testing problem, we use the test statistic

$$\frac{n\widehat{\sigma^2}}{\sigma^2} \sim \chi_n^2$$

Let $\chi_n^2(\alpha/2)$ and $\chi_n^2(1-\alpha/2)$ be the $(\alpha/2) \times 100$-th and $(1-\alpha/2) \times 100$-th percentiles, respectively. From the previous discussion, we know that we would accept H_0 if

$$\chi_n^2(\alpha/2) \leq \frac{n\widehat{\sigma^2}}{\sigma_0^2} \leq \chi_n^2(1-\alpha/2)$$

or

$$\frac{n\widehat{\sigma^2}}{\chi_n^2(1-\alpha/2)} \leq \sigma_0^2 \leq \frac{n\widehat{\sigma^2}}{\chi_n^2(\alpha/2)}$$

From this we can get the $1-\alpha$ confidence interval for σ^2 is

$$\left[\frac{n\widehat{\sigma^2}}{\chi_n^2(1-\alpha/2)}, \frac{n\widehat{\sigma^2}}{\chi_n^2(\alpha/2)}\right]$$

And the above result also says that if the value of σ_0^2 falls inside of the $1-\alpha$ confidence interval, we should accept H_0.

These two examples demonstrate the duality between confidence intervals and hypothesis testing.

6 Exercises

Exercise 1. We are interested in testing whether or not a coin is fair based on the number of heads X on 36 tosses of the coin.

a. Write out the null hypothesis and alternative hypothesis, figure out the testing statistic.

b. What is the null distribution of the test statistic?

c. Find the general form of the rejection region.

d. If the rejection region is $|X - 18| \geq 4$, calculate the value of α.

e. If the rejection region is $|X - 18| \geq 4$, calculate the value of β if the probability of head is 0.7.

Exercise 2. Suppose that $\mathbf{X} = (X_1, \cdots, X_n)$ is a random sample from a normal distribution with unknown mean μ and known variance σ^2. We wish to test the hypotheses

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_a : \mu > \mu_0.$$

at the significant level α. Calculate the probability of the type II error β, assuming the true value is μ_a, which satisfies $\mu_a > \mu_0$.

Exercise 3. Suppose that $\mathbf{X} = (X_1, \cdots, X_n)$ is a random sample from a normal distribution with unknown mean μ and known variance σ^2. We wish to test the hypotheses

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_a : \mu < \mu_0.$$

and we specify the value of desired α and β. β is evaluated when $\mu = \mu_a < \mu_0$, where μ_a is the true value of μ. Determine the sample size n.

Exercise 4. Suppose that $\mathbf{X} = (X_1, \cdots, X_n)$ is a random sample from a normal distribution with unknown mean μ and known variance σ^2. We wish to test the following hypotheses at the significance level α. Suppose the observed values are $x_1, \cdots, x_n$. For each case, find the expression of the p-value, and state your decision rule based on the p-values

a. $H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_a : \mu \neq \mu_0.$

b. $H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_a : \mu > \mu_0.$

Exercise 5. Suppose that the null hypothesis is true, that the distribution of the test statistic, say T, is continuous with cumulative distribution function F and that the test rejects the null hypothesis for large values of T. Let V denote the p-value of the test.

a. Show that $V = 1 - F(T)$.

b. Conclude that the null distribution of V is uniform.

c. If the null hypothesis is true, what is the probability that the p-value is greater than 0.1?

Exercise 6. Suppose $X_1, \cdots, X_n$ from a normal distribution $N(\mu, \sigma^2)$ where μ is unknown but σ is known. Consider the following hypothesis testing problem:

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_a : \mu > \mu_0$$

Prove that the decision rule is that we reject H_0 if

$$\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > Z(1 - \alpha),$$

where α is the significant level, and show that this is equivalent to rejecting H_0 if μ_0 is less than the $100(1 - \alpha)\%$ lower confidence bound for μ.