Hypotheses and Tests

Example

1. **(Fairness of Coin)** A person wanted to check if a coin was fair. He tossed the coin 20 times and got 13 heads. Is this evidence that the coin is not fair?

2. **(Production Procedure)** To test a manufacturing procedure, 500 items it produced were randomly sampled and 34 of them turned to be defective. Is this evidence that the procedure has a defective rate lower than 10%?

3. **(Filling Coke Bottles)** A machine at a Coke production plant is designed to fill bottles with 16 oz of Coke. The actual amount varies slightly from bottle to bottle. From past experience, it is known that the standard deviation of the amount is 0.2 oz. A random sample of 100 bottles filled by the machine has a mean 15.96 oz per bottle. Is this evidence that the machine needs to be re-calibrated?

Each of the problems involves a certain claim about a population whose plausibility needs to be investigated.

Hypothesis and Test. A **statistical hypothesis** is a claim about a population, whether it is about a single parameter, the values of several parameters, or the form an an entire probability distribution. A **hypothesis test** is an assessment of the evidence provided by a data set in favor of (or against) a hypothesis about a population.

1. In a hypothesis test, we want to decide if the *observed* value of a specific *sample* statistic is consistent with a hypothesis on the corresponding *population* parameter. The sample statistic is thus referred to as a *test statistic*.
2. If the observed value of the test statistic and the hypothesized value of the parameter differ (as they almost certainly will), we need to assess whether the difference is due to an incorrect hypothesis or is merely due to chance variation.

Null vs Alternative Hypotheses. The **null hypothesis**, denoted H_0, is the claim that is initially assumed to be true (the "prior belief" claim). The **alternative hypothesis**, denoted H_a, is the assertion that is contradictory to H_0.

Example (Cont'd) For each problem, the corresponding hypothesis can be formulated as follows.

1. $H_0 : p = 1/2$, $H_a : p \neq 1/2$, where p is the probability that the coin lands on head in a toss. Another reasonable alternative hypothesis is $H_a : p > 1/2$.
2. $H_0 : p = .1$, $H_a : p < .1$, where p is the defective rate of the procedure.
3. $H_0 : \mu = 16$, $H_a : \mu \neq 16$, where μ is the mean of the amount of Coke filled by the machine. Another reasonable alternative hypothesis is $H_a : \mu < 16$.

The null hypothesis will be rejected in favor of the alternative hypothesis only if sample evidence suggests that H_0 is false. If the sample does not strongly contradict H_0, we will continue to believe in the plausibility of H_0. The two possible conclusions from a hypothesis test are then

1. *reject* H_0, or
2. *fail to reject* H_0.

General Procedure for Hypothesis Testing

Before devising any test procedure, first, formulate H_0 and H_a carefully. This is the most important step of a scientific investigation.

1. H_0 is the statement being tested. In this course, it typically states a plausible value of the parameter under investigation, e.g., $\theta = \theta_0$, with the implication that the inconsistency between the observation and the hypothesized value is only due to chance variation. In the Coke example, H_0 is $\mu = 16$. The hypothesized value θ_0 is referred to as the **null value**.

2. H_a is the statement we will favor if we find evidence that H_0 is false. It usually states that there is a real inconsistency between the observation and the hypothesized value of the parameter. In the Coke example, H_a can be $\mu \neq 16$, or $\mu > 16$, or $\mu < 16$.

- H_a is **two-sided** if it has the form $H_a : \theta \neq \theta_0$.
- H_a is **one-sided** if it has the form $H_a : \theta > \theta_0$ or $H_a : \theta < \theta_0$.

After H_0 and H_a are formulated, a test procedure is specified by the following.

1. A **test statistic**, a function of the sample data on which the decision ("reject H_0 or not") is to be based. It should provide a measure of the difference between the observed data and what would be expected (e.g., null value) *if* H_0 is true.

2. A **rejection region**, the set of all test statistic values for which H_0 will be rejected. H_0 will then be rejected if and only if the observed or computed test statistic value falls in the rejection region.

The following outcomes are possible when conducting a test:

Reality	Test Result	
	H_0	H_a
H_0	$\surd$	**type I error**
H_a	**type II error**	$\surd$

For a real-world problem, it is rarely the case that an error-free procedure is available, as this demands an examination of the entire population. Instead, one should seek procedures that have small probabilities of errors. Traditionally, one denotes

$$\alpha = P(\text{type I error}) = P(\text{reject } H_0 \,|\, H_0 \text{ is true})$$

$$\beta = P(\text{type II error}) = P(\text{not reject } H_0 \,|\, H_a \text{ is true})$$

1. $1 - \beta$ is often referred to as the **power** of the test.

2. For a specific test on $H_0 : \theta = \theta_0$ *vs* $H_a : \theta \neq \theta_0$, or $H_0 : \theta = \theta_0$ *vs* $H_a : \theta < \theta_0$ (say), because H_0 specifies a unique value of θ, there is a unique value of α. However, there is a different value of β for each value of θ consistent with H_a.

Example In the coin toss example, the null hypothesis is

$$H_0 : p = 1/2$$

where p is the probability that the coin lands on head. We consider the *upper-tailed* alternative hypothesis

$$H_a : p > 1/2$$

The test statistic is

$$X = \text{number of heads in 20 tosses}$$

Suppose one sets the rejection region as

$$R = \{13, 14, 15, 16, 17, 18, 19, 20\}$$

The rejection region is *upper-tailed* because it consists only of large values of X.

When H_0 is true, $X \sim \text{Bin}(20, .5)$. Then

$$\begin{aligned}\alpha &= P(\text{type I error}) = P(H_0 \text{ is rejected when it is true}) \\ &= P(X \geq 13 \text{ when } X \sim \text{Bin}(20, .5)) \\ &= 1 - B(12; 20, .5) = 1 - .868 = .132\end{aligned}$$

That is, when H_0 is actually true (i.e., the coin is fair), roughly 13% of all experiments consisting of 20 tosses of the coin would result in the test incorrectly rejecting H_0 in favor of H_a.

In contrast to α, there is not a single β. Instead, there is a different β for each different $p > 1/2$. For example, if $p = .6$, then

$$\begin{aligned}\beta &= P(\text{type II error}) = P(H_0 \text{ not rejected when } p = 0.6) \\ &= P(X \leq 12 \text{ when } X \sim \text{Bin}(20, .6)) = B(12; 20, .6) = .584\end{aligned}$$

Thus, if p is actually .6 rather than .5, roughly 58% of all experiments consisting of 20 tosses of the coin would result in the test incorrectly not rejecting H_0 in favor of H_a.

Now suppose another person sets the rejection region as

$$R = \{15, 16, 17, 18, 19, 20\}$$

Comparing to the previous rejection region, this one makes it harder to reject H_0. Consequently, the probability of type I error should *decrease*, while the probability of type II error should *increase*. Indeed,

$$\begin{aligned}\alpha &= P(\text{type I error}) = P(H_0 \text{ is rejected when it is true}) \\ &= P(X \geq 15 \text{ when } X \sim \text{Bin}(20, .5)) \\ &= 1 - B(14; 20, .5) = 1 - .979 = .021\end{aligned}$$

and, if $p = .6$ as before,

$$\begin{aligned}\beta &= P(\text{type II error}) = P(H_0 \text{ not rejected when } p = 0.6) \\ &= P(X \leq 14 \text{ when } X \sim \text{Bin}(20, .6)) = B(14; 20, .6) = .874\end{aligned}$$ □

Example For the Coke example, assume the amount of Coke in a bottle has a normal distribution. Then the distribution of the amount of Coke in a bottle filled by the particular machine is $N(\mu, \sigma^2)$ with $\sigma = .2$. The null is

$$H_0 : \mu = 16$$

Suppose we determined beforehand that the alternative hypothesis should be

$$H_a : \mu \neq 16$$

Now if $\overline{X}$ is the sample mean of a random sample of 100 bottles filled by the machine, then $\overline{X}$ has distribution $N(\mu, \sigma^2/100)$.

Suppose the rejection region is set as

$$R = \{x : |x - 16| > .0392\} = (-\infty, 15.9608) \cup (16.0392, \infty)$$

Then

$$\begin{aligned}\alpha &= P(\overline{X} < 15.9608 \text{ or } \overline{X} > 16.0392 \text{ when } \overline{X} \sim N(16, .02^2)) \\ &= \Phi\left(\frac{15.9608 - 16}{.02}\right) + 1 - \Phi\left(\frac{16.0392 - 16}{.02}\right) \\ &= \Phi(-1.96) + 1 - \Phi(1.96) = .05\end{aligned}$$

On the other hand, if $\mu = 15.97$, then

$$\begin{aligned}\beta &= P(15.9608 \le \overline{X} \le 16.0392 \text{ when } X \sim N(15.97, .02^2)) \\ &= \Phi\left(\frac{16.0392 - 15.97}{.02}\right) - \Phi\left(\frac{15.9608 - 15.97}{.02}\right) = .677\end{aligned}$$

Now, if the rejection region is set as

$$R = \{x : |x - 16| > .05152\} = (-\infty, 15.94848) \cup (16.05152, \infty)$$

then

$$\begin{aligned}\alpha &= P(\overline{X} < 15.94848 \text{ or } \overline{X} > 16.05152 \text{ when } \overline{X} \sim N(16, .02^2)) \\ &= \Phi\left(\frac{15.94848 - 16}{.02}\right) + 1 - \Phi\left(\frac{16.05152 - 16}{.02}\right) \\ &= \Phi(-2.576) + 1 - \Phi(2.576) = .01\end{aligned}$$

and, again, if $\mu = 15.97$, then

$$\begin{aligned}\beta &= P(15.94848 \leq \overline{X} \leq 16.05152 \text{ when } X \sim N(15.97, .02^2)) \\ &= \Phi\left(\frac{16.05152 - 15.97}{.02}\right) - \Phi\left(\frac{15.94848 - 15.97}{.02}\right) = .859\end{aligned}$$

□

The examples illustrate an important fact about hypothesis testing: one has to make a trade-off between the probability of type I errors and the probability of type II errors.

> ***Trade-off between α and β.*** Suppose an experiment and a sample size are *fixed* and a test statistic is chosen. Then decreasing the rejection region results in a smaller value of α, while a larger value of β for any particular parameter value consistent with H_a.

The standard approach to type I error and type II error is as follows.

1. Specify the largest value of α that can be tolerated. This value is often referred to as the **significance level** of the test. Traditional levels of significance are .10, .05, and .01.

2. Identify a rejection region that has α no greater than the significance level and has the value of β as small as possible, or equivalently, is as powerful as possible.

Tests About a Population Mean

Suppose the unknown mean μ of a population is of interest. Suppose the null

$$H_0 : \mu = \mu_0$$

needs to be tested. Depending on the problem at hand, the alternative hypothesis typically is one of the following

$$\begin{aligned} H_a &: \mu > \mu_0 \quad \text{upper-tailed alternative} \\ H_a &: \mu < \mu_0 \quad \text{lower-tailed alternative} \\ H_a &: \mu \neq \mu_0 \quad \text{two-tailed alternative} \end{aligned}$$

To test the hypothesis, let $\widehat{\mu}$ be an unbiased point estimator of μ. The idea is to examine how different it is from μ_0. A suitable test statistic is the *z-score*

$$Z = \begin{cases} (\widehat{\mu} - \mu_0)/\sigma_{\widehat{\mu}} & \text{if the SD of } \widehat{\mu} \text{ is known} \\ (\widehat{\mu} - \mu_0)/\widehat{\sigma}_{\widehat{\mu}} & \text{otherwise.} \end{cases}$$

Next, we need to identify the rejection region for the test statistic Z.

1. For $H_a : \mu > \mu_0$, a rejection region consistent with the alternative should have the form $R = [c, \infty)$. In order to achieve significance level α, we need

$$P(Z \geq c \text{ when } H_0 \text{ is true}) = \alpha \text{ (or } \approx \alpha)$$

2. For $H_a : \mu < \mu_0$, a rejection region consistent with the alternative should have the form $R = (-\infty, c]$. In order to achieve significance level α, we need

$$P(Z \leq c \text{ when } H_0 \text{ is true}) = \alpha \text{ (or } \approx \alpha)$$

3. For $H_a : \mu \neq \mu_0$, a rejection region consistent with the alternative may have the form $R = \{z : |z| \geq c\}$ with $c > 0$. In order to achieve significance level α, we need

$$P(|Z| \geq c \text{ when } H_0 \text{ is true}) = \alpha \text{ (or } \approx \alpha)$$

The question is how to determine c. This will be considered in three cases.

Normal Population with Known σ^2. Let $\overline{X}$ be the mean of a random sample of size n from $N(\mu, \sigma^2)$. Then $\widehat{\mu} = \overline{X}$ is an unbiased estimator of μ and

$$Z = \frac{\widehat{\mu} - \mu}{\sigma_{\widehat{\mu}}} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

If H_0 is true, then $Z \sim N(0, 1)$ and hence

1. for $H_a : \mu > \mu_0$,

$$P(Z \geq c \text{ when } H_0 \text{ is true}) = 1 - \Phi(c)$$

So, to achieve significance level α, $c = z_\alpha$.

2. for $H_a : \mu < \mu_0$,

$$P(Z \leq c \text{ when } H_0 \text{ is true}) = \Phi(c)$$

So, to achieve significance level α, $c = -z_\alpha$.

3. for $H_a : \mu \neq \mu_0$,

$$P(|Z| \geq c \text{ when } H_0 \text{ is true}) = 2\Phi(c)$$

So, to achieve significance level α, $c = z_{\alpha/2}$.

***z* Test.** Let $\overline{x}$ be the observed mean value of a sample of size n from $N(\mu, \sigma^2)$ with σ^2 known. Let

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$$

To test $H_0 : \mu = \mu_0$ at significance level α, the rejection criterion is

H_a	criterion to reject H_0	equivalent criterion
$\mu > \mu_0$	$z \geq z_\alpha$	$\overline{x} \geq \mu_0 + z_\alpha \cdot \sigma/\sqrt{n}$
$\mu < \mu_0$	$z \leq -z_\alpha$	$\overline{x} \leq \mu_0 - z_\alpha \cdot \sigma/\sqrt{n}$
$\mu \neq \mu_0$	$\lvert z\rvert \geq z_{\alpha/2}$	$\lvert\overline{x} - \mu_0\rvert \geq z_{\alpha/2} \cdot \sigma/\sqrt{n}$

Example For the Coke example, $\mu_0 = 16$, $\sigma = 0.2$, $n = 100$ and $\overline{x} = 15.96$. Then

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{15.96 - 16}{0.2/\sqrt{100}} = -2.0$$

If we test $H_0 : \mu = \mu_0$ *vs* $H_a : \mu \neq \mu_0$ at significance level $\alpha = .05$, then $z_{\alpha/2} = 1.96$. Since $|z| = 2 > z_{\alpha/2}$, H_0 is rejected at significance level .05.

If we test $H_0 : \mu = \mu_0$ *vs* $H_a : \mu \neq \mu_0$ at significance level $\alpha = .01$, then $z_{\alpha/2} = 2.58$. Since $|z| < z_{\alpha/2}$, H_0 is not rejected in favor of H_a at significance level .01. □

Large-Sample Tests. Let $X_1, X_2, \ldots, X_n$ be a random sample from a non-normal population with unknown mean μ and variance. Then a test statistic for $H_0 : \mu = \mu_0$ is

$$Z = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$$

When n is sufficiently large, Z is approximately a standard normal rv. A rule of thumb

$$n > 40$$

will again be used. The rejection regions given in the previous case then result in significance level approximately α.

Large-Sample Approximate z Test. Let $\overline{x}$ be the observed mean value of a sample of size n from a population with unknown mean μ and variance. Let

$$z = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

To test $H_0 : \mu = \mu_0$ at significance level *approximate* α, the rejection criterion is

H_a	criterion to reject H_0	equivalent criterion
$\mu > \mu_0$	$z \geq z_\alpha$	$\overline{x} \geq \mu_0 + z_\alpha \cdot s/\sqrt{n}$
$\mu < \mu_0$	$z \leq -z_\alpha$	$\overline{x} \leq \mu_0 - z_\alpha \cdot s/\sqrt{n}$
$\mu \neq \mu_0$	$\lvert z \rvert \geq z_{\alpha/2}$	$\lvert \overline{x} - \mu_0 \rvert \geq z_{\alpha/2} \cdot s/\sqrt{n}$

Example To test

$$H_0 : \mu = 30 \quad vs. \quad H_a : \mu < 30$$

at significance level .05, a sample of 52 observations was collected to get $\overline{x} = 28.76$, $s^2 = 12.26^2$. In this case $\mu_0 = 30$ and

$$z = \frac{\overline{x} - \mu_0}{s/\sqrt{52}} = -.73$$

Since $z_\alpha = z_{.05} = 1.65$, we find $z > -z_\alpha$. Therefore H_0 cannot be rejected in favor of H_a at significance level .05. □

Normal Population with Unknown σ^2. As in the previous case, a test statistic for $H_0 : \mu = \mu_0$ is $(\overline{X} - \mu_0)/(S/\sqrt{n})$. If H_0 is true, then the test statistic is known to be t_{n-1}. Thus denote

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

Note that T is derived from a single sample. Such tests are often referred to as one-sample tests.

One-Sample t Test. Let $\overline{x}$ be the observed mean value of a sample of size n from a normal population with unknown mean μ and variance. Let

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

To test $H_0 : \mu = \mu_0$ at significance level *approximate* α, the rejection criterion is

H_a	criterion to reject H_0	equivalent criterion
$\mu > \mu_0$	$t \geq t_{\alpha,n-1}$	$\overline{x} \geq \mu_0 + t_{\alpha,n-1} \cdot s/\sqrt{n}$
$\mu < \mu_0$	$t \leq -t_{\alpha,n-1}$	$\overline{x} \leq \mu_0 - t_{\alpha,n-1} \cdot s/\sqrt{n}$
$\mu \neq \mu_0$	$\lvert t\rvert \geq t_{\alpha/2,n-1}$	$\lvert\overline{x} - \mu_0\rvert \geq t_{\alpha/2,n-1} \cdot s/\sqrt{n}$

Example To test

$$H_0 : \mu = 4 \quad vs. \quad H_a : \mu \neq 4$$

at significance level 5% for a normal population with σ^2 unknown, a sample of 5 observations was collected to get $\overline{x} = 3.814$ and $s^2 = .718^2$. In this case $\mu_0 = 4$. Because the sample size is small, we should use t test to get

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = -.58$$

Since $t_{\alpha/2,n-1} = t_{.025,4} = 2.776$, we find $|t| < t_{\alpha/2,n-1}$. Therefore, H_0 cannot be rejected in favor of H_a as significance level 5%. □

Tests on Population Proportion

Suppose for a particular population, the proportion of individuals possessing a specified property is p. To test

$$H_0 : p = p_0,$$

let n individuals be randomly sampled from the population and let X be the number of individuals in the sample that possess the property. Then $X \sim \text{Bin}(n, p)$. If n is large, then, since $\widehat{p} = X/n$ is an unbiased estimator of p, one can use the standardized $\widehat{p} = X/n$ as a test statistic and use large-sample approximate z-test. If n is small, then one can directly use X as a test statistic and use the Binomial table to test H_0.

Large-sample tests. When n is large, in principle one still can use the binomial table to test H_0. However, it is much more convenient and quite accurate to use large-sample approximation.

Like all the other large-sample tests, the test statistic is $\widehat{p} = X/n$ standardized in some way. Since $\sigma_{\widehat{p}} = \sqrt{p(1-p)/n}$ and p is unknown, $\sigma_{\widehat{p}}$ is unknown. However, if H_0 is true, then

$$\sigma_{\widehat{p}} = \sqrt{p_0(1-p_0)/n}$$

and in the large-sample case,

$$Z = \frac{\widehat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

is approximately a standard normal rv.

Large-Sample Approximate z Test for Population Proportion. Suppose the null is

$$H_0 : p = p_0$$

Let

$$z = \frac{\widehat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

Then at significance level *approximate* α, the rejection criterion is

H_a	criterion to reject H_0
$p > p_0$	$z \geq z_\alpha$
$p < p_0$	$z \leq -z_\alpha$
$p \neq p_0$	$\lvert z \rvert \geq z_{\alpha/2}$

These test procedures are valid provided that $np_0 \geq 10$ and $n(1 - p_0) \geq 10$.

Example Let p be the defective rate of a certain product. To test

$$H_0 : p = .15 \quad vs. \quad H_a : p > .15$$

at significance level 10%, a sample of size 91 was collected with 16 of the sampled items being defective. In this case, $p_0 = .15$. Since $np_0 > 10$ and $n(1-p_0) > 10$, the large-sample z test can be used to get

$$z = \frac{\widehat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{16/91 - .15}{\sqrt{(.15)(.85)/91}} = .69$$

Since $z_\alpha = z_{.10} = 1.28$, we get $z < z_\alpha$. Therefore H_0 cannot be rejected in favor of H_a as significance level 10%. □

Small-sample tests. If n is small, X can be used as a test statistic.

To test

$$H_0 : p = p_0 \quad vs \quad H_a : p > p_0$$

as seen previously, the rejection region is $R = \{c, c+1, \ldots, n\}$ for a suitable c, such that H_0 is rejected if and only if $X \geq c$. By

$$\begin{aligned} P(\text{type I error}) &= P(H_0 \text{ is rejected when it is true}) \\ &= P(X \geq c \text{ when } X \sim \text{Bin}(p_0, n)) = 1 - B(c-1; n, p_0) \end{aligned}$$

to achieve significance level α, there must be $1 - B(c-1; n, p_0) \leq \alpha$. Then, to achieve the minimum probability of type II error, c should be the *smallest* integer satisfying

$$B(c-1; n, p_0) \geq 1 - \alpha$$

If $p = p' > p_0$, the probability of type II error is then

$$P(\text{type II error}) = P(X < c \text{ when } X \sim \text{Bin}(n, p')) = B(c-1; n, p')$$

Example For the Coin example, the null is $H_0 : p = 1/2$, where p is the probability that the coin lands on head. To test H_0 *vs* $H_a : p > 1/2$ at significance level .05, c is the smallest integer satisfying

$$B(c-1; 20, .5) \geq .95$$

From the binomial table, $B(13; 20, .5) = .942$, $B(14; 20, .5) = .979$, so $c = 15$. Since the observed number of heads is $13 < 15$, H_0 is not rejected in favor of H_a at significance level .05.

If in reality $p = p' > 1/2$, it is clear that the larger the difference $p' - 1/2$, the more likely the above test rejects H_0. What is the smallest value of p' in order for the test to be able to reject H_0 in favor of H_a with probability at least .7? As $c = 15$, the value is the smallest p' satisfying

$$B(14; 20, p') \leq 1 - .7 = .3$$

From the binomial table, $B(14; 20, .75) = .383$ and $B(14; 20, .8) = .196$. Therefore, the smallest value is between .75 and .8. Numerical calculation shows it is about .771. □

Similarly, to test

$$H_0 : p = p_0 \ vs \ H_a : p < p_0$$

a parallel argument can be used. At significance level α, the rejection region is $R = \{0, 1, \ldots, c\}$, where c is the *largest* number satisfying

$$B(c; n, p_0) \leq \alpha$$

To test

$$H_0 : p = p_0 \ vs \ H_a : p \neq p_0$$

the situation is a little different since two cut-off values are needed. At significance level α, the rejection region is $R = \{0, 1, \ldots, c_1\} \cup \{c_2, c_2 + 1, \ldots, n\}$, where c_1 and c_2 are two numbers such that

$$B(c_1; n, p_0) + 1 - B(c_2 - 1; n, p_0) \leq \alpha$$

and such that R contains as many numbers as possible.

P-values

Let Z be a test statistic to test a null hypothesis H_0. If z is the observed value of Z in an experiment, then its ***P*-value** is the probability of Z being *at least as extreme as* z assuming the null hypothesis H_0 is true.

In a test on the mean μ of a population, if the null is $H_0 : \mu = \mu_0$ and Z is a standardized sample mean $\overline{X}$, then the P-value is as follows.

H_a	P-value of z
$\mu > \mu_0$	$P(Z \geq z \text{ if } H_0 \text{ is true})$
$\mu < \mu_0$	$P(Z \leq z \text{ if } H_0 \text{ is true})$
$\mu \neq \mu_0$	$P(\lvert Z\rvert \geq \lvert z\rvert \text{ if } H_0 \text{ is true})$

The P-value is similarly evaluated for tests on a population proportion. For the large-sample test, the P-value is defined the same as above. For the small-sample test,

H_a	P-value of x
$p > p_0$	$P(X \geq x$ if H_0 is true$)$
$p < p_0$	$P(X \leq x$ if H_0 is true$)$

Some comments on P-values.

1. The smaller the P-value, the stronger the evidence *against* H_0.
2. H_0 is rejected at significance level α if and only if P-value $\leq \alpha$.
3. For different observed values of the test statistic, the P-value is different.
4. The P-value is the smallest significance level α at which H_0 can be rejected. Thus the P-value is also referred to as the **observed significance level (OSL)** for the data.
5. The P-value is *not* the probability that H_0 is true, nor is it an error probability!

A traditional rough interpretation of the P-value is as follows

P-value	Interpretation
$P > .1$	no evidence against H_0
$.05 < P \leq .1$	weak evidence against H_0
$.01 < P \leq .05$	evidence against H_0
$P \leq .01$	strong evidence against H_0

P-Values for z Tests. To test $H_0 : \mu = \mu_0$ for a population,

$$Z = \begin{cases} \dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} & \text{if population is normal, } \sigma^2 \text{ known} \\ \dfrac{\overline{X} - \mu_0}{S/\sqrt{n}} & \text{if population is not normal} \end{cases}$$

can be used as a test statistic. In the first case, since $Z \sim N(0, 1)$, the P-value of an observed value z of Z is

$$P = \begin{cases} 1 - \Phi(z) & \text{for } z \text{ test on } H_0 \textit{ vs } H_a : \mu > \mu_0 \\ \Phi(z) & \text{for } z \text{ test on } H_0 \textit{ vs } H_a : \mu < \mu_0 \\ 2[1 - \Phi(|z|)] & \text{for } z \text{ test on } H_0 \textit{ vs } H_a : \mu \neq \mu_0 \end{cases}$$

In the second case, as long as Z is sufficiently large, each value is an *approximate* P-value of z for the corresponding test.

Example In the Coke example, we got $z = -2.0$. Since the test is two-tailed, the P-value of z is

$$P = 2[1 - \Phi(|-2.0|)] = (2) \cdot (.0228) = .0456$$

As a result, we again conclude that H_0 is rejected in favor of H_a at significance level 5%, but not at significance level 1%. □

***P*-Values for *t* Tests.** With a parallel argument, for a normal population with unknown σ^2, the P-value of an observed value t of the T statistic is

$$\begin{cases} 1 - T(t; n-1) & \text{for } t \text{ test on } H_0 \text{ } vs \text{ } H_a : \mu > \mu_0 \\ T(t; n-1) & \text{for } t \text{ test on } H_0 \text{ } vs \text{ } H_a : \mu < \mu_0 \\ 2[1 - T(|t|; n-1)] & \text{for } t \text{ test on } H_0 \text{ } vs \text{ } H_a : \mu \neq \mu_0 \end{cases}$$

where $T(t; \nu)$ denotes the cdf of T_ν, i.e., $T(t; \nu)$ is the area under the t_ν density curve to the left of t.

Statistical vs. practical significance.

Saying that a result is *statistically significant* does not signify that it is large or necessarily important. That decision depends on the particulars of the problem. A statistically significant result only says that there is substantial evidence that H_0 is false.

Failure to reject H_0 does not imply that H_0 is correct. It only implies that *we have insufficient evidence to conclude that* H_0 *is incorrect.*

Potential abuses of tests.

In many applications, a researcher tests a null hypothesis with the intent of discrediting it. For example:

- H_0: new drug has the same effect as placebo
- H_0: men and women are paid equally

A small P-value can help a drug company to get a drug approved by the FDA. Similarly, a researcher may have an easier time publishing his results if the P-value is smaller than .05.

Because of that we have to be aware of the following potential abuses.

1. Using one-sided tests to make the P-value one-half as big
2. Conducting repeated sampling and testing and reporting only the lowest P-value
3. Testing many hypotheses or testing the same hypothesis on many different subgroups.

In the last two, even if there is actually no effect, you will probably get at least one small P-value.