Hypothesis Testing

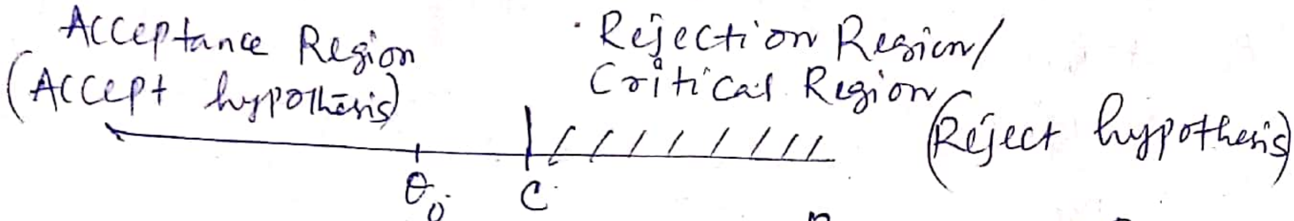
A statistical hypothesis is an assumption about the distribution of a random variables. A statistical test of a hypothesis is a procedure in which a sample is used to find out whether we may not reject the hypothesis, that is, act as though it is true, or whether we should reject it, that is act as though it is false.

The hypothesis to be tested is sometimes called the null hypothesis and a counter assumption is called an alternative hypothesis. The number α is called the significance level of the test, c is called the critical value. The region containing the values for which we reject the hypothesis is called the rejection region or critical region. The region of values for which we do not reject the hypothesis is called the acceptance region.

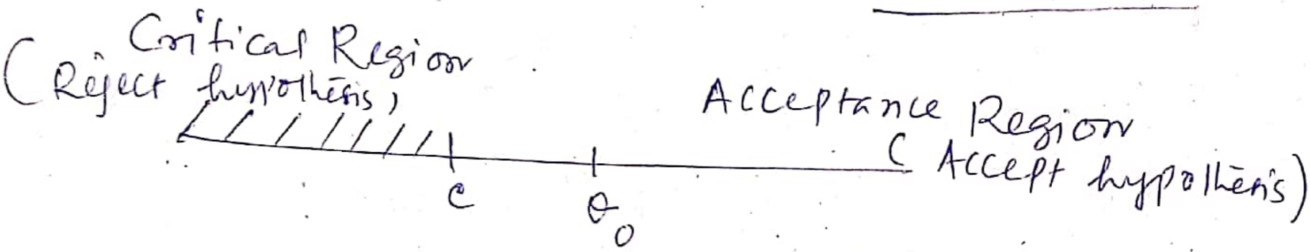
Let θ be the unknown parameter in a distribution and suppose that we want to test the hypothesis $\theta = \theta_0$. There are three main types of alternatives,

$$\left.\begin{array}{l} \theta > \theta_0 \\ \theta < \theta_0 \end{array}\right\} \text{One-sided alternatives}$$

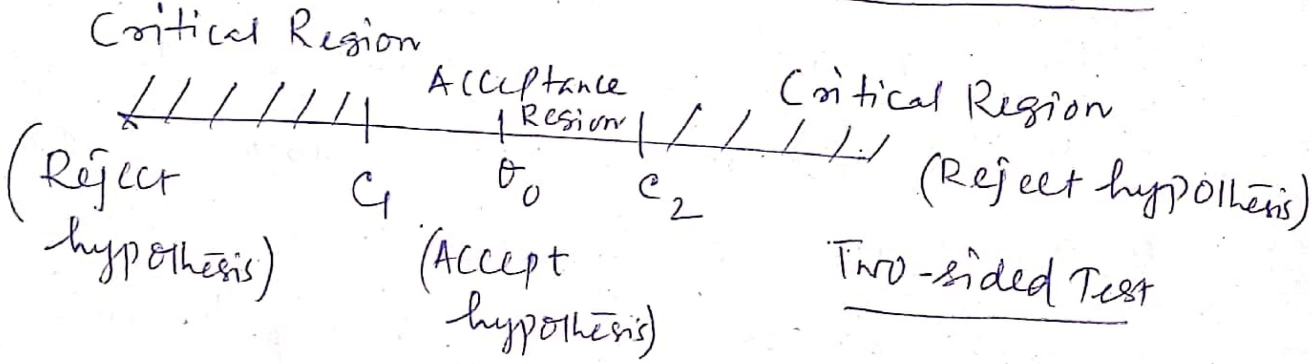
$$\theta \neq \theta_0 \rightarrow \text{two-sided alternative.}$$



Right-sided Test



Left-sided Test



Two-sided Test

Types of Error in Hypothesis Testing:

Type I Error: The hypothesis is true, but it is rejected.

Type II Error: The hypothesis is false, but it is accepted

Type I and Type II Errors in testing a hypothesis $\theta = \theta_0$ against an alternative $\theta = \theta_1$

	Unknown Truth $\theta = \theta_0$	$\theta = \theta_1$
Accepted $\theta = \theta_0$	True decision $P = 1 - \alpha$	Type II Error $P = \beta$
$\theta = \theta_1$	Type I Error $P = \alpha$	True decision $P = 1 - \beta$

Note $\eta = 1 - \beta$ is called power of the test.

Q. A firm sells oil in cans containing 1000g oil per can and is interested to know whether the mean weight differs significantly from 1000g at the 5% level, in which case the filling machine has to be adjusted. Set up a hypothesis and an alternative and perform the test, assuming normality and using a sample of 20 fillings having a mean of 996g and a standard deviation of 5g.

Sol^n. $H_0 : \mu = 1000$ (Null hypothesis)

$H_1 ; \mu \neq 1000$ (Alternate hypothesis)

Given $\bar{x} = 996$, $s = 5$, $n = 20$, $\alpha = 5\%$

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \sqrt{n}\left(\frac{\bar{x} - \mu_0}{s}\right) = \frac{\sqrt{20}\,(996 - 1000)}{5}$$

$$= -3.58$$

Form t-distribution table with 19 degrees of freedom

$$C = -2.09.$$

As $t < C$

Reject the hypothesis.

Scanned with CamScanner

Approach to Hypothesis Testing with fixed probability of Type I Error:

1. State the null and alternate hypotheses.
2. Choose a fixed significance level α.
3. Choose an appropriate test statistics and establish the critical region based on α.
4. Reject H_0 if the computed test statistic is in the critical region.
   Otherwise do not reject.
5. Draw Conclusion.

Q. A random sample of 100 recorded deaths in USA during the past year showed an average life span of 71.8 years. Assuming a population standard deviation of 8.9 years, does this seem to indicate that the mean life span today is greater than 70 years? Use a 0.05 level of significance.

Soln $H_0: \mu = 70$ years.

$H_1: \mu > 70$ years

$\alpha = 0.05$

Critical region $z > 1.645$

$$\text{Where } z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{71.8 - 70}{8.9/\sqrt{100}} = 2.02$$

$\therefore$ Reject H_0

$\therefore$ Mean life span today is greater than 70 years.

Q. A manufacturer of sports equipment has developed a new synthetic fishing line that the company claims has a mean breaking strength of 8 kg with a standard deviation 0.5 kg. Test the hypothesis $\mu = 8$ kg against the alternative that $\mu \neq 8$ kg if a random sample of 50 lines is tested and found to have a mean breaking strength of 7.8 kg. Use 0.01 level of significance.

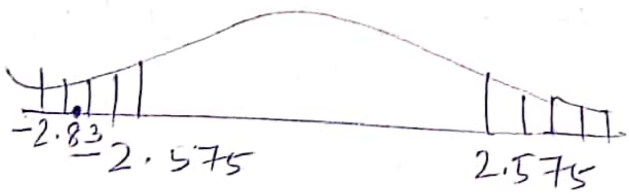
Sol^n: $H_0: \mu = 8$ kg

$H_1: \mu \neq 8$ kg

$\alpha = 0.01$.

Critical region $Z < -2.575$ and $Z > 2.575$

where $Z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \dfrac{7.8 - 8}{0.5/\sqrt{50}} = -2.83$



Reject H_0.

$\therefore$ The average breaking strength is not equal to 8kg but in fact less than 8 kg.

Q. A vacuum cleaner uses an average of 46 kilowatt hours per year. If a random sample of 12 homes included in a planned study indicates that vacuum cleaner use an average of 42 kilowatt hours per year with a standard deviation of 11.9 kilowatt hours, does this suggest at the 0.05 level of significance that vacuum cleaner use, on average less than 46 kilowatt hours annually? Assume the population of kilowatt hours to be normal.

Solution: $H_0: \mu_0 = 46$ kilowatt hours

$H_1: \mu_1 < 46$ kilowatt hours

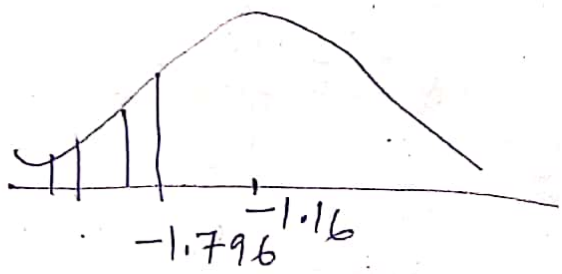
Given $\alpha = 0.05$

Critical region $t < -1.796$ with 11 degrees of freedom.

$$\text{where } t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{42-46}{11.9/\sqrt{12}} = -1.16$$

Do not reject H_0.

Hence the conclusion is that the average number of kilowatt hours used annually by home vacuum cleaner is not significantly less than 46 kilowatt hours.



23.20 Pairs of Measurements. Fitting Straight Lines

We shall now discuss experiments in which we observe or measure two quantities simultaneously. In practice we may distinguish between two types of experiments, as follows.

1. In **correlation analysis** both quantities are random variables and we are interested in relations between them. (We shall not discuss this branch of statistics.)
2. In **regression analysis** one of the two variables, call it x, can be regarded as an ordinary variable, that is, can be measured without appreciable error. The other variable, Y, is a random variable. x is called the *independent* (sometimes the *controlled*) *variable,* and one is interested in the dependence of Y on x. Typical examples are the dependence of the blood pressure Y on the age x of a person or, as we shall now say, the regression of Y on x, the regression of the gain of weight Y of certain animals on the daily ration of food x, the regression of the heat conductivity Y of cork on the specific weight x of the cork, etc.

In the experiment the experimenter first selects n values $x_1, \cdots, x_n$ of x and then observes Y at those values of x, so that he obtains a sample of the form $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$. In regression analysis the mean μ of Y is assumed to depend on x, that is, is a function $\mu = \mu(x)$ in the ordinary sense. The curve of $\mu(x)$ is called the *regression curve of Y on x*. In the present section we shall discuss the simplest case, when $\mu(x)$ is a linear function, $\mu(x) = \alpha + \beta x$. Then we may want to plot the sample values as n points in the xY-plane, fit a straight line through them, and use this line for estimating $\mu(x)$ for given values of x. so that we know what values of Y we can expect if we choose certain values of x. If the points are scattered, fitting "by eye" becomes unreliable and we need a mathematical method for fitting lines that yields a unique result depending only on the points. A widely used procedure is the **method of least squares** developed by Gauss. In our present situation it may be formulated as follows.

The straight line should be fitted through the given points so that the sum of the squares of the distances of those points from the straight line is minimum, where the distance is measured in the vertical direction (the y-direction).

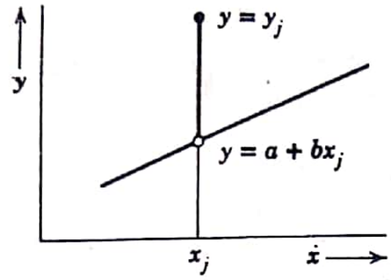


Fig. 455. Vertical distance of a point (x_j, y_j) from a line $y = a + bx$

General assumption (A1)

The x-values $x_1, \cdots, x_n$ of our sample $(x_1, y_1), \cdots, (x_n, y_n)$ are not all equal.

Consider a sample $(x_1, y_1), \cdots, (x_n, y_n)$ of size n. The vertical distance (distance measured in the y-direction) of a sample value (x_j, y_j) from a straight line $y = a + bx$ is $|y_j - a - bx_j|$; cf. Fig. 455. Hence the sum of the squares of these distances is

(1) $$q = \sum_{j=1}^{n} (y_j - a - bx_j)^2.$$

In the method of least squares we choose a and b such that q is minimum. q depends on a and b, and a necessary condition for q to be minimum is

(2) $$\frac{\partial q}{\partial a} = 0 \quad \text{and} \quad \frac{\partial q}{\partial b} = 0.$$

We shall see that from this condition we obtain the formula

(3) $$y - \bar{y} = b(x - \bar{x})$$

where

(4) $$\bar{x} = \frac{1}{n}(x_1 + \cdots + x_n) \quad \text{and} \quad \bar{y} = \frac{1}{n}(y_1 + \cdots + y_n).$$

(3) is called the **regression line** *of the y-values of the sample on the x-values of the sample.* Its slope b is called the **regression coefficient** *of y on x,* and we shall see that

(5) $$b = \frac{s_{xy}}{s_1^2}.$$

Here,

(6) $$s_1^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})^2 = \frac{1}{n-1}\left[\sum_{j=1}^{n} x_j^2 - \frac{1}{n}\left(\sum_{j=1}^{n} x_j\right)^2\right]$$

and

(7) $$s_{xy} = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y}) = \frac{1}{n-1}\left[\sum_{j=1}^{n} x_j y_j - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{j=1}^{n} y_j\right)\right].$$

s_{xy} is called the **covariance** of the sample. Obviously, the regression line (3) passes through the point $(\bar{x}, \bar{y})$.

To derive (3), we use (1) and (2), finding

$$\frac{\partial q}{\partial a} = -2 \sum (y_j - a - bx_j) = 0$$

$$\frac{\partial q}{\partial b} = -2 \sum x_j(y_j - a - bx_j) = 0$$

(where we sum over j from 1 to n). Thus

$$na + b \sum x_j = \sum y_j$$

$$a \sum x_j + b \sum x_j^2 = \sum x_j y_j.$$

Because of Assumption (A1), the determinant

$$n \sum x_j^2 - \left(\sum x_j\right)^2 = n(n-1)s_1^2$$

[cf. (6)] of this system of linear equations is not zero, and the system has a unique solution [cf. (4), (6), (7)]

$$\text{(8)} \qquad a = \bar{y} - b\bar{x}, \qquad b = \frac{n \sum x_j y_j - \sum x_j \sum y_j}{n(n-1)s_1^2}.$$

This yields (3) with b given by (5)–(7). (The equality of the two expressions for s_1^2 in (6) may be shown by the reader (cf. Prob. 13); similarly for (7).)

Hand calculations can be simplified by *coding*, that is, by setting

$$\text{(9)} \qquad x_j = c_1 x_j^* + l_1, \qquad y_j = c_2 y_j^* + l_2$$

and choosing the constants c_1, c_2, l_1, l_2 such that the transformed values x_j^* and y_j^* are as simple as possible. We first compute the values $\bar{x}^*$, $\bar{y}^*$, x_1^{*2}, s_{xy}^* corresponding to the transformed values and then

$$\text{(10)} \qquad \begin{aligned} \bar{x} &= c_1 \bar{x}^* + l_1, & \bar{y} &= c_2 \bar{y}^* + l_2 \\ s_1^2 &= c_1^2 s_1^{*2}, & s_{xy} &= c_1 c_2 s_{xy}^*. \end{aligned}$$

Q. Find a regression line of y on x of the following dat

x	6	9	11	13	22	26	28	33	35
y	68	67	65	53	44	40	37	34	32

Solⁿ: The regression line of y on x is

$$y - \bar{y} = b(x - \bar{x}) \quad \text{—(1)}$$

Where b is called the regression coefficient of y on x

and $b = \dfrac{S_{xy}}{S_1^2}$

where $S_1^2 = \dfrac{1}{n-1} \sum_{j=1}^{n} (x_j - \bar{x})^2$

and $S_{xy} = \dfrac{1}{n-1} \sum_{j=1}^{n} (x_j - \bar{x})(y_j - \bar{y})$

Here $n = 9$, $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{9}[6+9+11+13+22+26+28+33+35]$

$= 20.33$

$\bar{y} = \frac{1}{n} \sum_{i=1}^{} y_i = \frac{1}{9}[68+67+65+53+44+40+37+34+32]$

$= 48.889$

$$S_1^2 = \frac{1}{9-1} \sum_{j=1}^{9} (x_j - 20.33)^2$$

$$= \frac{1}{8}\left[(6-20.33)^2 + (9-20.33)^2 + (11-20.33)^2 + (13-20.33)^2 + (22-20.33)^2 + (26-20.33)^2 + (28-20.33)^2 + (33-20.33)^2 + (35-20.33)^2\right]$$

$$\Rightarrow s_1^2 = \frac{1}{8}[205.3489 + 128.3689 + 87.0489 + 53.7289 + 2.7889 + 32.1489 + 58.8289 + 160.5289 + 215.2089]$$

$$= \frac{944.0001}{8} = 118.0000125$$

$$S_{xy} = \frac{1}{9-1}\sum_{j=1}^{9}(x_j - 20.33)(y_j - 48.89)$$

$$= \frac{1}{8}[(6-20.33)(68-48.89) + (9-20.33)(67-48.89) + (11-20.33)(65-48.89) + (13-20.33)(53-48.89) + (22-20.33)(44-48.89) + (26-20.33)(40-48.89) + (28-20.33)(37-48.89) + (33-20.33)(34-48.89) + (35-20.33)(32-48.89)]$$

$$= \frac{1}{8}[(-14.33)(19.11) + (-11.33)(18.11) + (-9.33)(16.11) + (-7.33)(4.11) + (1.67)(-4.89) + (5.67)(-8.89) + (7.67)(-11.89) + (12.67)(-14.89) + (14.67)(-16.89)]$$

$$= \frac{1}{8}[-273.8463 - 205.1863 - 150.3063 - 30.1263 - 8.1663 - 50.4063 - 91.1963 - 188.6563 - 247.7763]$$

$$= \frac{1}{8}[-1245.6667] = -155.7083375$$

$$\therefore b = \frac{-155.7083375}{118.0000125} = -1.3195 \approx -1.32$$

$\therefore$ From (1) the required regression line of y on x is

$$\boxed{y - 48.89 = -1.32(x - 20.33)}$$

Scanned with CamScanner