



Z Test for 0-1 Box Proportion

Hypothesis

Null hypothesis H_0 : $p = p_0$, the unknown proportion p is equal to the special value p_0 .

Two-tailed alternative hypothesis H_1 : $p \neq p_0$.

Right tail alternative hypothesis H_1 : $p > p_0$.

Left tail alternative hypothesis H_1 : $p < p_0$.

Assumptions

The data comes from a random sample of a 0-1 box.

Test Statistic

$$Z = \frac{\bar{X} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}$$

P-value

Two-sided test:

```
2 * pnorm(abs(stat), lower.tail = F)
```

Right tail test:

```
pnorm(stat, lower.tail = F)
```

Left tail test:

```
pnorm(stat, lower.tail = T)
```

Z Test

Hypothesis

Null hypothesis H_0 : $\mu = \mu_0$, the unknown proportion p is equal to the special value p_0 .

Two-tailed alternative hypothesis H_1 : $\mu \neq \mu_0$.

Right tail alternative hypothesis H_1 : $\mu > \mu_0$.

Left tail alternative hypothesis H_1 : $\mu < \mu_0$.

Assumptions

Data is normally distributed or CLT is applicable. The data comes from a random sample of the population.

Test Statistic

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

P-value

Two-sided test:

```
2 * pnorm(abs(stat), lower.tail = F)
```

Right tail test:

```
pnorm(stat, lower.tail = F)
```

Left tail test:

```
pnorm(stat, lower.tail = T)
```

T Test

Hypothesis

Null hypothesis H_0 : $\mu = \mu_0$, the unknown proportion p is equal to the special value p_0 .

Two-tailed alternative hypothesis H_1 : $\mu \neq \mu_0$.

Right tail alternative hypothesis H_1 : $\mu > \mu_0$.

Left tail alternative hypothesis H_1 : $\mu < \mu_0$.

Assumptions

Data is normally distributed or CLT is applicable. The data comes from a random sample of the population.

Test Statistic

$$T = \frac{\bar{X} - \mu_0}{\frac{\hat{\sigma}}{\sqrt{n}}} \sim t_{n-1}$$

where

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

P-value

Two-sided test:

```
2 * pt(abs(stat), df = n - 1, lower.tail = F)
```

Right tail test:

```
pt(stat, df = n - 1, lower.tail = F)
```

Left tail test:

```
pt(stat, df = n - 1, lower.tail = T)
```

Bootstrap simulation

```
1. // Method 1
2. stat = (mean(sample_data) - mu0) / (sd(sample_data) / sqrt(n))
3. sim.stat = 0
4. sample_data.g = sample_data - mean(sample_data) + mu0
5.
6. for (i in 1:num_sim) {
7.     boot_sample = sample(sample_data.g, size = n, replace = TRUE)
8.     sim.stat[i] = (mean(boot_sample) - mu0) / (sd(boot_sample) / sqrt(n))
9. }
10.
11. p_value = mean(abs(sim.stat) >= abs(stat))
```

```
1. // Method 2
2. stat = (mean(sample_data) - mu0) / (sd(sample_data) / sqrt(n))
3. sim.stat = 0
4.
5. for (i in 1:num_sim) {
6.     boot_sample = sample(sample_data, size = n, replace = TRUE)
7.     sim.stat[i] = (mean(boot_sample) - mean(sample_data)) / (sd(boot_sample) / sqrt(n))
8. }
```

```
8. }  
9.  
10. p_value = mean(abs(sim.stat) >= abs(stat))
```

Paired T Test

Hypothesis

Null hypothesis H_0 : $\mu_X = \mu_Y$ or $\mu_{diff} = 0$.

Two-tailed alternative hypothesis H_1 : $\mu_X \neq \mu_Y$ or $\mu_{diff} \neq 0$.

Right tail alternative hypothesis H_1 : $\mu_X > \mu_Y$ or $\mu_{diff} > 0$.

Left tail alternative hypothesis H_1 : $\mu_X < \mu_Y$ or $\mu_{diff} < 0$.

Assumptions

Each of the paired measurements must be obtained from the same subject. Each pair comes from a random sample of the population. The differences are normally distributed or CLT is applicable. (If both data set are approximately normal, then the difference is approximately normal as well.)

Test Statistic

Perform T test on the sample differences (D).

$$T = \frac{\bar{D}}{\frac{\hat{\sigma}}{\sqrt{n}}} \sim t_{n-1}$$

where

$$\hat{\sigma} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (D_i - \bar{D})^2}$$

P-value

Two-sided test:

```
1. // Method 1  
2. t.test(X - Y)  
3. // Method 2  
4. t.test(X, Y, paired = T)
```

```
5. // Method 3
6. 2 * pt(abs(stat), df = n - 1, lower.tail = F)
```

Right tail test:

```
1. // Method 1
2. t.test(X - Y, alternative = "greater")
3. // Method 2
4. t.test(X, Y, paired = T, alternative = "greater")
5. // Method 2
6. pt(stat, df = n - 1, lower.tail = F)
```

Left tail test:

```
1. // Method 1
2. t.test(X - Y, alternative = "less")
3. // Method 2
4. t.test(X, Y, paired = T, alternative = "less ")
5. // Method 2
6. pt(stat, df = n - 1, lower.tail = T)
```

Bootstrap Simulation

```
1. diff = X - Y
2. stat = (mean(diff)) / (sd(diff) / sqrt(n))
3. sim.stat = 0
4.
5. for(i in 1:num_sim) {
6.     boot.samp = sample(diff, size = n, replace = T)
7.     sim.stat[i] = (mean(boot.samp) - mean(diff)) / (sd(boot.samp) / sqrt(n))
8. }
9.
10. p_value = mean(abs(sim.stat) >= abs(stat))
```

Two-Sample Z Test

Hypothesis

Null hypothesis H_0 : $\mu_X = \mu_Y$ or $\mu_{diff} = 0$.

Two-tailed alternative hypothesis H_1 : $\mu_X \neq \mu_Y$ or $\mu_{diff} \neq 0$.

Assumptions

Both data set are normally distributed or CLT is applicable. Data set X comes from a random sample of population X, data set Y sample comes from a random sample of population Y.

Test Statistic

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}} \sim N(0,1)$$

P-value

```
2 * pnorm(abs(stat), lower.tail = F)
```

Two-Sample T Test

Hypothesis

Null hypothesis H_0 : $\mu_X = \mu_Y$ or $\mu_{diff} = 0$.

Two-tailed alternative hypothesis H_1 : $\mu_X \neq \mu_Y$ or $\mu_{diff} \neq 0$.

Assumptions

Both data set are normally distributed or CLT is applicable. Data set X comes from a random sample of population X, data set Y sample comes from a random sample of population Y. The variances within the two groups should be roughly equal.

Test Statistic

$$T = \frac{\bar{X} - \bar{Y}}{\hat{\sigma}_p \sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t_{m+n-2}$$

where

$$\hat{\sigma}_p = \sqrt{\frac{\sum_{i=1}^m (X_i - \bar{X})^2 + \sum_{j=1}^n (Y_j - \bar{Y})^2}{m+n-2}} = \sqrt{\frac{(m-1)\hat{\sigma}_X^2 + (n-1)\hat{\sigma}_Y^2}{m+n-2}}$$

is called pooled estimate of σ (weighted average of $\hat{\sigma}_X^2$ and $\hat{\sigma}_Y^2$).

P-value

```
2 * pt(abs(stat), df = n - 1, lower.tail = F)
```

Bootstrap Simulation

```
1. pooled_sd = sqrt(((m - 1) * sd(X)^2 + (n - 1) * sd(Y)^2) / (m + n - 2))
2. stat = (mean(X) - mean(Y)) / (pooled_sd * sqrt(1/m + 1/n))
3. sim.stat = 0
4. X.g = X - mean(X)
5. Y.g = Y - mean(Y)
6.
7. for (i in 1:num_sim) {
8.   boot.x = sample(X.g, size = m, replace = TRUE)
9.   boot.y = sample(Y.g, size = n, replace = TRUE)
10.  pooled_sd = sqrt(((m-1) * sd(boot.x)^2 + (n-1) * sd(boot.y)^2) / (m + n - 2))
11.  sim.stat[i] = (mean(boot.x) - mean(boot.y)) / (pooled_sd * sqrt(1/m + 1/n))
12. }
13.
14. p_value <- mean(abs(sim.stat) >= abs(stat))
```

Welch's T Test

Hypothesis

Null hypothesis H_0 : $\mu_X = \mu_Y$ or $\mu_{diff} = 0$.

Two-tailed alternative hypothesis H_1 : $\mu_X \neq \mu_Y$ or $\mu_{diff} \neq 0$.

Assumptions

Both data set are normally distributed or CLT is applicable. Data set X comes from a random sample of population X, data set Y sample comes from a random sample of population Y.

Test Statistic

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\hat{\sigma}_X^2}{m} + \frac{\hat{\sigma}_Y^2}{n}}} \sim t_{dof}$$

where the degrees of freedom is a complicated function of m , n , σ_X and σ_Y .

P-value

Welch's T Test is the default two-sample T test in R.

```
t.test(X, Y)
```

Bootstrap simulation

```
1. est.SE = sqrt((sd(X) ^ 2) / m + (sd(Y) ^ 2) / n)
2. stat = (mean(X) - mean(Y)) / est.SE
3.
4. X.g = X-mean(X)
5. Y.g = Y-mean(Y)
6. stat.sim = 0
7.
8. for(i in 1:num_sim) {
9.     boot.x = sample(X.g, size = m, replace = T)
10.    boot.y = sample(Y.g, size = n, replace = T)
11.    boot.SE = sqrt((sd(boot.x) ^ 2) / m + (sd(boot.y) ^ 2) / n)
12.    stat.sim[i] = (mean(boot.x) - mean(boot.y)) / boot.SE
13. }
14.
15. p_value = mean(abs(stat.sim) >= abs(stat))
```

Chi-Squared Goodness of Fit Test

Hypothesis

Null hypothesis H_0 : $\mathbf{p} = \mathbf{p}_0$ for some hypothesized $\mathbf{p}_0 = (p_{01}, \dots, p_{0k})$.

Alternative hypothesis H_1 : $\exists i$ such that $p_i \neq p_{0i}$.

Assumptions

The population is assumed to be divided into k categories. The data comes from a random sample of the population. The expected frequencies of each category should be at least 5.

Test Statistic

$$T = \sum_{j=1}^k \frac{(O_j - E_j)^2}{E_j}$$
$$T \overset{approx.}{\sim} \chi_{k-1}^2$$

where O_j is the number of data points labelled j and E_j is the expected frequencies under H_0 , $E_j = np_{0j}$.

P-value

```
1. // Method 1
2. pchisq(stat, df = k - 1, lower.tail=F)
3. // Method 2
4. chisq.test(O, p = p0)
```

Simulation

```
1. Oi = tabulate(samp, nbins = k) # works even if some values don't appear
2. stat = chisq.test(Oi, p = p0)$stat
3. // Method 1
4. sim.stat = 0 # the dice example
5. for(i in 1:100000) {
6.     sim.rolls = sample(box, size = n, replace = T)
7.     freqs = tabulate(sim.rolls, nbins=k) # works even with zero freqs, better
than table()
8.     sim.stat[i] = chisq.test(freqs)$stat # save the test statistics
9. }
10. mean(sim.stat >= stat)
11. // Method 2
12. chisq.test(Obs.freq, p = p0, simulate = T, B = 100000)
```

Chi-Squared Test of Independence

Hypothesis

Null hypothesis: the events {being in Row i } and {being in Col j } are independent. That is H_0 : $p_{ij} = P\{\text{in Row } i \text{ and Col } j\} = P\{\text{in Row } i\} \times P\{\text{in Col } j\} = p_{i.}p_{.j}$.

Alternative hypothesis H_1 : $p_{ij} \neq p_{i.}p_{.j}$.

Assumptions

The data comes from a random sample of the population. Each observation belongs to a unique combination of categories. The expected frequencies of each category combination should be at least 5.

Test Statistic

$$T = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$$T^{approx.} \sim \chi_{(r-1)(c-1)}^2$$

where O_{ij} is the observed frequencies and E_{ij} is the expected frequencies under H_0 ,

$$E_{ij} = np_{i.}p_{.j} = \frac{o_{i.}o_{.j}}{n}.$$

	Col 1	Col 2	...	Col c	Total
Row 1	$np_{1.}p_{.1}$	$np_{1.}p_{.2}$...	$np_{1.}p_{.c}$	$np_{1.}$
Row 2	$np_{2.}p_{.1}$	$np_{2.}p_{.2}$...	$np_{2.}p_{.c}$	$np_{2.}$
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
Row r	$np_{r.}p_{.1}$	$np_{r.}p_{.2}$...	$np_{r.}p_{.c}$	$np_{r.}$
Total	$np_{.1}$	$np_{.2}$...	$np_{.c}$	n

P-value

```
1. // Method 1
2. pchisq(stat, df = d, lower.tail = F)
3. // Method 2
4. chisq.test(Oij)
```

Simulation

```
chisq.test(Oij, simulate = T)
```

T Test for Slope

Simple Linear Regression

Correlation coefficient:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

Regression line connects (\bar{x}, \bar{y}) to $(\bar{x} + SD_x, \bar{y} + r \cdot SD_y)$.

Simple linear regression model:

$$Y_i = b_0 + b_1 x_{1i} + \varepsilon_i$$

where

$$b_1 = r \cdot \frac{SD_y}{SD_x}$$

$$b_0 = \bar{y} - b_1 \cdot \bar{x}_1$$

$$\varepsilon_i(b_0, b_1) = y_i - \hat{y}_i = y_i - (b_0 + b_1 \cdot x_{1i})$$

Hypothesis

Null Hypothesis H_0 : $b_1 = 0$ there is no linear relationship between x_1 and Y .

Two-tailed alternative hypothesis H_1 : $b_1 \neq 0$.

Right tail alternative hypothesis H_1 : $b_1 > 0$.

Left tail alternative hypothesis H_1 : $b_1 < 0$.

Assumptions

The model is $Y_i = b_0 + b_1 x_{1i} + \varepsilon_i$, where $\varepsilon_i \sim (iid) N(0, \sigma^2)$. “iid” stands for independent and identically distributed and σ is the SD of the error box.

Test Statistic

$$T = \frac{\hat{b}_1 - b_1}{\widehat{SE}(\hat{b}_1)} = \frac{\hat{b}_1}{\widehat{SE}(\hat{b}_1)} \sim t_{n-2}$$

where

$$\begin{aligned} \widehat{SE}(\hat{b}_j) &= \frac{\hat{\sigma}}{\sqrt{SST \text{ in } x_1}} = \sqrt{\frac{1}{n - (p + 1)} \frac{SSE}{SST \text{ in } x_1}} \\ &= \sqrt{\frac{1}{n - (p + 1)} \frac{\sum_{i=1}^n (y_i - (\hat{b}_0 + \hat{b}_1 x_{1i}))^2}{\sum_{i=1}^n (x_{1i} - \bar{x}_1)^2}} \end{aligned}$$

P-value

```
summary(model) # where model = lm (y ~ x1)
```

```

1. Call:
2. lm(formula = y ~ x1)
3.
4. Residuals:
5.      Min       1Q   Median       3Q      Max
6. -8.8772 -1.5144 -0.0079  1.6285  8.9685
7.
8. Coefficients:
9.              Estimate Std. Error t value Pr(>|t|)
10. (Intercept)  33.88660    1.83235   18.49  <2e-16 ***
11. x1           0.51409     0.02705   19.01  <2e-16 ***
12. ---
13. Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
14.
15. Residual standard error: 2.437 on 1076 degrees of freedom
16. Multiple R-squared:  0.2513,    Adjusted R-squared:  0.2506
17. F-statistic: 361.2 on 1 and 1076 DF,  p-value: < 2.2e-16

```

T Test for Individual Coefficient

Multiple Linear Regression

If we have multiple independent variables x_1, x_2, \dots, x_p , the linear model becomes

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x_1 + \hat{b}_2 x_2 + \dots + \hat{b}_p x_p$$

$$Y_i = \hat{b}_0 + \hat{b}_1 x_{1i} + \hat{b}_2 x_{2i} + \dots + \hat{b}_p x_{pi} + \varepsilon_i$$

$$Y = \beta X + \varepsilon$$

where

$$Y = (Y_1, Y_2, \dots, Y_n)'$$

$$\beta = (b_0, b_1, \dots, b_p)'$$

$$X = \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{p1} \\ 1 & x_{12} & x_{22} & \dots & x_{p2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \dots & x_{pn} \end{bmatrix}$$

$$\varepsilon \sim N_n(\mathbf{0}, \sigma^2 I)$$

Hypothesis

Null Hypothesis $H_0: b_j = 0$ there is no linear relationship between x_j and Y .

Two-tailed alternative hypothesis $H_1: b_j \neq 0$.

Right tail alternative hypothesis $H_1: b_j > 0$.

Left tail alternative hypothesis $H_1: b_j < 0$.

Assumptions

The model is $\mathbf{Y} = \boldsymbol{\beta}\mathbf{X} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim (iid) N_n(\mathbf{0}, \sigma^2 \mathbf{I})$. No linear relationship exists between independent variables.

Test Statistic

$$T = \frac{\hat{b}_j - b_j}{\widehat{SE}(\hat{b}_j)} = \frac{\hat{b}_j}{\widehat{SE}(\hat{b}_j)} \sim t_{n-(p+1)}$$

where

$$\widehat{SE}(\hat{b}_j) = \hat{\sigma} \times \sqrt{[(\mathbf{X}'\mathbf{X})^{-1}]_{jj}}$$
$$\hat{\sigma} = \sqrt{\frac{SSE}{n - (p + 1)}}$$

P-value

```
summary(lm3)
```

```
1. Call:
2. lm(formula = log.ozone ~ radiation + temperature + wind, data = env.new)
3.
4. Residuals:
5.      Min       1Q   Median       3Q      Max
6. -2.06212 -0.29968 -0.00223  0.30767  1.23572
7.
8. Coefficients:
9.              Estimate Std. Error t value Pr(>|t|)
10. (Intercept) -0.2611739  0.5534102  -0.472 0.637934
11. radiation    0.0025147  0.0005567   4.518 1.62e-05 ***
```

```

12. temperature  0.0491630  0.0060863   8.078 1.07e-12 ***
13. wind         -0.0615925  0.0157037  -3.922 0.000155 ***
14. ---
15. Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
16.
17. Residual standard error: 0.5085 on 107 degrees of freedom
18. Multiple R-squared:  0.6645,    Adjusted R-squared:  0.6551
19. F-statistic: 70.65 on 3 and 107 DF,  p-value: < 2.2e-16

```

Partial and Overall F Test

Hypothesis

Partial F-test null hypothesis $H_0: b_1 = b_2 = 0$, **some** regression coefficients (except the intercept) are zero. The additional independent variables (x_1 and x_2 in this example) have no effect in explaining Y . That is: $Y_i = b_0 + b_3x_{3i} + b_4x_{4i} + \dots + b_px_{pi} + \varepsilon_i$.

Overall F-test null hypothesis $H_0: b_1 = b_2 = \dots = b_p = 0$, **all** regression coefficients (except the intercept) are zero. That is: $Y_i = b_0 + \varepsilon_i$.

Alternative hypothesis $H_1: \exists b_j \neq 0$, at least one of the regression coefficients is not zero.

Assumptions

The model is $Y = \beta X + \varepsilon$, where $\varepsilon \sim (iid) N_n(\mathbf{0}, \sigma^2 I)$. No linear relationship exists between independent variables.

Test Statistic

Consider a null model with q independent variables and an alternative model with p independent variables. The alternative model is always larger, so $p > q$.

Under H_0 : Fit the model and calculate \widehat{SSE}_{H_0} . Degrees of freedom is $n - (q + 1)$

Under H_1 : Fit the model and calculate \widehat{SSE}_{H_1} . Degrees of freedom is $n - (p + 1)$

$$F = \frac{(\widehat{SSE}_{H_0} - \widehat{SSE}_{H_1}) / (p - q)}{\widehat{SSE}_{H_1} / (n - (p + 1))} \sim F_{p-q, n-(p+1)}$$

P-value

```
1. // Method 1
2. pf(stat, p - q, n - (p + 1), lower.tail = F)
3. // Method 2
4. summary(lm3)
```

Adjusted R-squared

Adjusted R-squared penalizes the inclusion of unhelpful independent variables.

Adjusted R-squared

$$= 1 - \frac{\text{Estimated SD of the residual error}}{\text{Sample SD of the dependent variable}}$$

$$= 1 - \frac{\hat{\sigma}}{\hat{s}_X}$$

$$= 1 - \frac{\widehat{SSE}/(n - (p + 1))}{\widehat{SST}/(n - 1)}$$

$$= 1 - (1 - r^2) \frac{n - 1}{n - (p + 1)}$$

$$\geq r^2$$