Statistical tests (proportions and categorical variables)

sabato 25 ottobre 2025 14:31

# One-sample proportion z-test

### 1. Purpose and Context

The **one-sample proportion z-test** is used to determine whether the **observed proportion of "successes"** in a sample differs significantly from a **hypothesized population proportion**.

It applies to **binary (yes/no, success/failure)** data — typical in quality control, medical testing, survey analysis, and similar fields.

Examples:

- Testing if the proportion of defective items differs from 5%.
- Checking if a candidate's support rate differs from 50%.
- Evaluating if a treatment's success rate is higher than a baseline.

### 2. Hypotheses

Let

- $p_0$: hypothesized population proportion,
- $\hat{p} = \frac{x}{n}$: sample proportion of successes, $\longrightarrow$ $\hat{p}$ will be called $Y$
- $n$: sample size.

Then we test:

$$H_0 : p = p_0$$

$$H_1 : p \neq p_0 \quad \text{(two-tailed)}$$

or one-tailed forms:

$$H_1 : p > p_0 \quad \text{or} \quad H_1 : p < p_0$$

### 3. Assumptions

1. The sample is **random** and consists of **independent Bernoulli trials** (each with the same success probability).
2. The **sample size is large enough** that both $np_0$ and $n(1-p_0)$ are ≥ 5.
   This ensures that the **sampling distribution of** $\hat{p}$ is approximately normal. $\longrightarrow$ $\hat{p}$ is $Y$

$$X_1 \dots X_n \sim Be(p) \text{ iid} \Rightarrow X := \sum_{i=1}^{n} X_i \sim Bi(n, p)$$

$$\Rightarrow Y := \frac{X}{n} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

for n large, a binomial distribution can be approximated by a normal distribution.

PROOF

$$*) \ \mathbb{E}(Y) = \mathbb{E}\left(\frac{X}{n}\right) = \frac{1}{n}\mathbb{E}\left(\sum X_i\right) = \frac{1}{n}\sum \mathbb{E}(X_i) = \frac{1}{\cancel{n}} \cancel{n} \cdot p = p$$

$$*) \ Var(Y) = Var\left(\frac{1}{n}\sum X_i\right) \underbrace{=}_{\text{for the independence of } X_i \, \forall i} \frac{1}{n^2}\sum Var(X_i) = \frac{1}{n^2} \cdot \cancel{n} \, p(1-p) = \frac{1}{n}p(1-p) \quad \square$$

$$\Rightarrow \quad Z := \frac{Y - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$$

$\Rightarrow$ Then is $Z$-test. So the decision rule (on a $\alpha$-level) will be to reject ~~$H_0$~~ if:

- for a two-tailed test $|Z| > z_{\alpha/2}$
- for a right test $Z > z_\alpha$
- for a left test $Z < -z_\alpha$

example

a quality inspector checks $n = 100$ items and finds $x = 8$ defective ones. We test whether the defect rate differs from 5%. at the 5% significance level.

$$Y = \frac{X}{n} = \frac{8}{100} = 0.08 \qquad p_0 = 0.05$$

$$\Rightarrow \quad Z = \frac{0.08 - 0.05}{\sqrt{\frac{0.05(1-0.05)}{100}}} \cong 1.376 \qquad \text{while} \quad z_{0.05} = \pm 1.96$$

$\Rightarrow$ Since $|1.376| < 1.96$, we fail to reject $H_0$. There is no significance evidence that the defect rate is different from 5%.

# Two-sample z-test for proportions

### 1. Purpose and Context

The **two-sample z-test for proportions** is used to test whether **two independent populations** have the **same proportion of success**.

It's widely applied in:

- comparing treatment success rates in two groups,
- testing differences in conversion rates between two web designs (A/B testing),
- comparing proportions of voters supporting two candidates, etc.

### 2. Hypotheses

Let

- $p_1, p_2$: true population proportions of success,
- $Y_1 = \dfrac{X_1}{n_1}, Y_2 = \dfrac{X_2}{n_2}$: observed sample proportions in the two groups.

We want to test whether the two population proportions are equal:

$$H_0 : p_1 = p_2$$

$$H_1 : p_1 \neq p_2 \quad \text{(two-tailed)}$$

or, for one-tailed alternatives:

$$H_1 : p_1 > p_2 \quad \text{or} \quad H_1 : p_1 < p_2$$

### 3. Assumptions

1. The two samples are **independent** and each consists of **Bernoulli trials** with success probabilities $p_1$ and $p_2$.
2. Under the null hypothesis, the two populations share a common success probability $p$.
3. The **sample sizes** are large enough for the normal approximation to hold, i.e.:

$$n_1 Y_1,\ n_1(1 - Y_1),\ n_2 Y_2,\ n_2(1 - Y_2) \geq 5$$

so that $Y_1$ and $Y_2$ are approximately normally distributed.

$X^{(1)} : X_1 \dots X_{n_1} \sim Be(p_1)$

$X^{(2)} : X_1 \dots X_{n_2} \sim Be(p_2)$

$Y_1 = \dfrac{X^{(1)}}{n_1} \qquad Y_2 = \dfrac{X^{(2)}}{n_2}$

$$\Rightarrow \hat{p} := \frac{X^{(1)} + X^{(2)}}{n_1 + n_2}$$

$$\Rightarrow z := \frac{(Y_1 - Y_2)}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

Can you guess the decision rule?

example

a marketing team is testing two ad campaigns

| GROUP | SAMPLE SIZE | SUCCESSES |
| --- | --- | --- |
| A | 200 | 60 |
| B | 250 | 90 |

$\Rightarrow \; y_1 = \frac{60}{200} = 0.30 \quad y_2 = \frac{90}{250} = 0.36$ and $\hat{p} = \frac{60+90}{200+250} = 0.33$

$$\Rightarrow \; SE = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} = \ldots = 0.0447$$

$$\Rightarrow \; z = \frac{0.30-0.36}{0.0447} = -1.34$$

$\Rightarrow \; |-1.34| < z_{0.05} = 1.96 \Rightarrow$ We fail to reject $H_0$

$\Rightarrow$ No significant evidence that the two campaigns have different conversion rates.

# Chi-square Goodness-of-Fit test

### 1. Purpose and Context

The **chi-square goodness-of-fit test** evaluates whether observed categorical frequencies across $k$ mutually exclusive categories are consistent with a specified theoretical distribution (multinomial with probabilities $p_1^0, \dots, p_k^0$).

Common uses:

- Checking whether a die is fair (each face with probability $1/6$).
- Verifying if observed genotype counts match Mendelian ratios.
- Testing whether observed market shares match an expected model.

This is a large-sample, frequency-based test that compares observed counts to expected counts under the null hypothesis.

### 2. Hypotheses

Let $O_i$ be the observed count in category $i$ (for $i = 1, \dots, k$) from a sample of size $n$. Let $p_i^0$ be the hypothesized probability for category $i$ under $H_0$.

The hypotheses are:

$$H_0: (p_1, p_2, \dots, p_k) = (p_1^0, p_2^0, \dots, p_k^0)$$

$$H_1: \text{the distribution of categories differs from the specified } (p_1^0, \dots, p_k^0)$$

(One may formulate directional or more specific alternatives, but the standard test is two-sided: any deviation from the specified vector.)

### 3. Assumptions

1. Observations form a **single multinomial sample** of size $n$ across the $k$ categories (i.e., each observation falls in exactly one category).
2. Expected counts under $H_0$, $E_i = np_i^0$, should be **sufficiently large** (common rule: $E_i \geq 5$ for all $i$; if not, combine categories or use exact/simulation methods).
3. Observations are **independent**.
4. The hypothesized probabilities $p_i^0$ are fully specified **a priori**. (If parameters are estimated from data, the degrees of freedom must be adjusted — see §6.)

Suppose to have k bins and for each bin i we observe $O_i$ frequencies, each of them with $p_i$ probability (i.e. the probability to fall into the $i^{th}$ bin is $p_i$). Under the null hypothesis, such probability is $p_i^0$ and therefore the expected value for each bin is $E_i$.

The Pearson chi-square statistic is:

$$\chi^2 := \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \xrightarrow[n \to +\infty]{d} \chi^2(k-1)$$

where $E_i = np_i^0$

The proof of this statement uses the Cochran theorem and the central limit theorem.

example

Test whether a six-sided die is fair using $n=120$ rolls with observed face counts:

$$O = (18, 22, 20, 17, 23, 20)$$

$p_i^0 = \frac{1}{6}$ so expected counts : $E_i = \frac{120}{6} = 20 \quad \forall i$

$$\Rightarrow \chi^2 = \frac{18-20}{20} + \frac{22-20}{20} + \cancel{\frac{20-20}{20}} + \frac{17-20}{20} + \frac{23-20}{20} + \cancel{\frac{20-20}{20}} = 1.3$$

and $\chi^2_{1-0.05}(6-1) = \chi^2_{0.95}(5) \cong 11.07$

$\Rightarrow$ Since $1.30 < 11.07$, we fail to reject $H_0$.
There is no significant evidence to suppose the die to be unfair.

# Chi-square test of indipendence

### 1. Purpose and Context

The **Chi-Square Test of Independence** determines whether two **categorical variables** are **statistically independent** or associated with one another.
It is among the most widely used nonparametric tests for categorical data analysis, particularly in **contingency tables**.

Common applications include:

- testing whether gender and voting preference are independent,
- checking if education level is related to smoking status,
- verifying whether product choice depends on region.

It is sometimes called the **test of association** or **contingency test**.

### 2. Hypotheses

Suppose we have a contingency table with:

- $r$ rows (categories of variable A),
- $c$ columns (categories of variable B),
- $O_{ij}$: observed count in cell $(i, j)$,
- $n = \sum_i \sum_j O_{ij}$: total sample size.

Then:

$$H_0 : \text{Variables A and B are independent}$$

$$H_1 : \text{Variables A and B are not independent (there is association)}$$

Under $H_0$, the expected cell counts are the products of the marginal proportions:

$$E_{ij} = \frac{(\text{row total})_i \times (\text{column total})_j}{n}$$

### 3. Assumptions

1. **Independence** of observations (each subject contributes to exactly one cell).
2. **Sample size large enough** so that all expected counts satisfy $E_{ij} \geq 5$ (a common rule of thumb).
3. Data come from **random sampling** or a randomized experiment.
4. Both variables are **categorical** (nominal or ordinal with few levels).

The Pearson statistic

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

example

a researcher studies whether smoking status is related to gender in a survey of $n=200$ people.

| | Smoker | No Smoker | Total row |
|---|---|---|---|
| Male | 40 | 60 | 100 |
| Female | 20 | 80 | 100 |
| Total column | 60 | 140 | 200 |

$$E_{11} = \frac{100 \cdot 60}{200} = 30 \qquad E_{12} = \frac{100 \cdot 140}{200} = 70$$

$$E_{21} = \frac{100 \cdot 60}{200} = 30 \qquad E_{22} = \frac{100 \cdot 140}{200} = 70$$

$$\Rightarrow \chi^2 = \frac{(40-30)^2}{30} + \frac{(60-70)^2}{70} + \frac{(20-30)^2}{30} + \frac{(80-70)^2}{70} = 9.52$$

while $\chi^2_{(2-1)(2-1)} {}_{0.95} = \chi^2_{(1)} {}_{0.95} = 3.84$

$\Rightarrow$ Since $9.52 > 3.84$ we reject $H_0$. There is statistically significant association between gender and smoking status.

# Fisher exact test

### 1. Purpose and Context

The **Fisher's Exact Test** evaluates whether two **categorical variables** (each with two levels) are **independent**, based on data from a **2×2 contingency table**.
It is especially useful when **sample sizes are small** or when some expected frequencies are **less than 5**, making the Chi-Square Test unreliable.

Common examples:

- Testing whether treatment and recovery are independent in a small clinical study.
- Examining whether gender is related to a binary outcome when sample sizes are small.

Unlike the chi-square test, Fisher's test computes the **exact probability** of observing the given table (or a more extreme one) under the null hypothesis, **without relying on large-sample approximations**.

### 2. Hypotheses

Suppose we have the following $2 \times 2$ table:

| | Outcome 1 | Outcome 2 | Row total |
|---|---|---|---|
| **Group 1** | $a$ | $b$ | $a+b$ |
| **Group 2** | $c$ | $d$ | $c+d$ |
| **Column total** | $a+c$ | $b+d$ | $n=a+b+c+d$ |

Then:

$$H_0 : \text{Variables are independent}$$

$$H_1 : \text{Variables are not independent (association exists)}$$

### 3. Assumptions

1. The data are counts from **two independent random samples** (or from one random sample classified by two binary variables).
2. The **marginal totals** (row and column totals) are **fixed** by design or conditioning.
3. Observations are **independent** within and between samples.
4. The table represents **dichotomous variables** (2×2).

If $H_0$ is true, then, given fixed marginals, the cell count a follows a hypergeometric (see pdf "Probability distributions") distribution:

$$\mathbb{P}(A=a) = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}}$$

$\rightsquigarrow$ For a two-tailed test, we compute the probability of all tables with the same marginals whose probabilities are less than or equal to that of the observed table (i.e. as or more extreme).

$\Rightarrow$ Let p-value be the sum of hypergeometric probabilities of all "extreme" tables.

At significance level $\alpha$:

- we reject $H_0$ if p-value $< \alpha$
- otherwise we fail to reject $H_0$.

example

Consider a small medical trial:

| | RECOVERED | NOT RECOVERED | TOTAL |
|---|---|---|---|
| TREATMENT | 4 (a) | 1 (b) | 5 |
| CONTROL | 1 (c) | 4 (d) | 5 |
| TOTAL | 5 | 5 | 10 (n) |

$$\mathbb{P}(A=4) = \frac{\binom{5}{4}\binom{5}{1}}{\binom{10}{5}} = 0.0992$$

The only other value of A that gives an equal or smaller probability is A=0 ( $\mathbb{P}(A=0) = 0.0992$ )

$\Rightarrow$ $p\text{-value} = \mathbb{P}(A=0) + \mathbb{P}(A=4) = 2 \cdot 0.0992 \simeq 0.1984 > \alpha = 0.05$

$\Rightarrow$ we fail to reject $H_0$. There is no statistically significant evidence to conclude that there is dipendence between the variables.