CI and z-Test for Population Proportion

As reported by AMA, 16% doctors in any given year are subject to malpractice claims. A hospital of 300 physicians received claims against 58 of their doctors.

Since $\hat{p} = 58/300 = .1933$, $\hat{\sigma}_{\hat{p}} = \sqrt{.1933(1-.1933)/300} = .0228$, a 95% CI is given by

$$.1933 \pm 1.96(.0228),$$

or $(0.1486, 0.2380)$.

To test for $H_0 : p = .16$,

$$z = \frac{\hat{p}-.16}{\sqrt{.16(1-.16)/300}} = 1.575,$$

with a 2-sided p-value of 0.115.

Consider $X \sim \text{Bin}(n,p)$. For n large, by CLT,

$$P\left(\frac{X/n - p}{\sqrt{p(1-p)/n}} \leq z\right) \approx \Phi(z).$$

The sample proportion $\hat{p} = X/n$ is actually an $\bar{X}$. As an estimate of $\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$, one may use

$$\hat{\sigma}_{\hat{p}} = \sqrt{\hat{p}(1-\hat{p})/n}.$$

A $(1-\alpha)100\%$ CI for p is thus

$$\hat{p} \pm z_{\alpha/2}\hat{\sigma}_{\hat{p}}.$$

To test $H_0 : p = p_0$ vs. $H_a : p \neq p_0$,

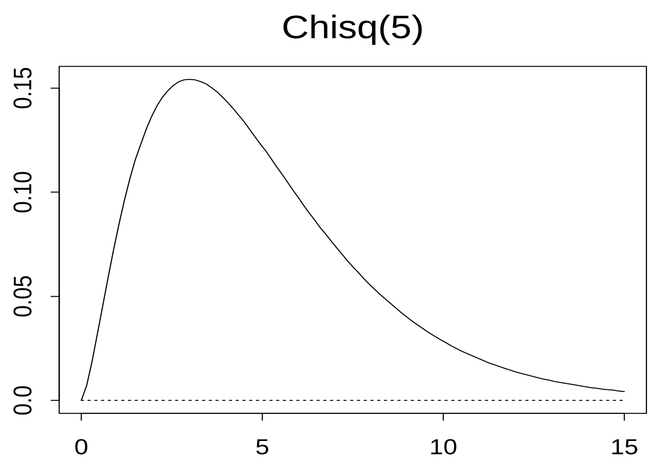
$$Z = \frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}}.$$

χ^2-Test for Binary Proportions

Let $Z_i \sim N(0,1)$, $i = 1, \ldots, n$, independent. The distribution of

$$\sum_{i=1}^n Z_i^2$$

is a χ^2**-distribution** with n degrees of freedom. Selected percentiles of χ^2-distributions, $\chi^2_{\alpha,\nu}$, can be found in Table A.11.



The z-test is equivalent to a χ^2-test based on the **expected and observed cell counts**.

E	np_0	$n(1-p_0)$
O	Y	$n-Y$

One rejects H_0 if

$$\chi^2 = \sum \frac{(O-E)^2}{E} \geq \chi^2_{\alpha,1}.$$

For the malpractice data,

E	48	252
O	58	242

$$\chi^2 = \frac{(58-48)^2}{48} + \frac{(252-242)^2}{252} = 2.480$$

- $2.480 = (1.575)^2$, $\chi^2_{\alpha,1} = 1.96^2$.

Testing for Multiple Cell Proportions

According to a certain Mendelian genetic model, self-pollination of pink-flowered plants of snapdragon should produce progeny that are red, pink, and white in the ratio 1:2:1.

$$H_0\colon p_1 = .25,\ p_2 = .5,\ p_3 = .25.$$

Data were obtained on 209 plants.

E	52.25	104.5	52.25
O	52	128	29

$$\chi^2 = \frac{.25^2}{52.25} + \frac{23.5^2}{104.5} + \frac{23.25^2}{52.25} = 15.63$$

Since $P(\chi^2_2 > 15.63) = .0004$, evidence was strong against H_0.

The χ^2-test applies to many testing problems involving cell proportions. To test for *given* proportions of multiple, say k, cells

$$H_0\colon p_i = p_{i0} \quad \text{vs.} \quad H_a\colon \text{o.w.},$$

one obtains the expected and observed cell counts,

E	np_{10}	$\cdots$	np_{k0}
O	Y_1	$\cdots$	Y_k

and rejects H_0 if

$$\chi^2 = \sum \frac{(O-E)^2}{E} \geq \chi^2_{\alpha,k-1}.$$

Note that $\sum Y_i = n$ and $\sum p_i = 1$. H_0 has 0 df; H_a has $k-1$.

χ^2-Test for Composite Hypothesis

2 alleles at a locus yield 3 blood types, MM, MN, and NN. In equilibrium, the 3 types should have probabilities θ^2, $2\theta(1-\theta)$, and $(1-\theta)^2$, respectively, where θ is the prevalence of M in the population. A sample of size 500 gives 125:225:150.

Based on $\hat{\theta} = (2(125)+225)/1000 = .475$, the estimated expected $\hat{E}$'s are 112.8:249.4:137.8. One has

$$\chi^2 = \frac{12.2^2}{112.8} + \frac{24.4^2}{249.4} + \frac{12.2^2}{137.8} = 4.787$$

Since $P(\chi^2_1 > 4.787) = .0287$, evidence was moderately strong against equilibrium.

When H_0 is not completely specified but pending on knowledge of some parameter(s), say

$$H_0 : p_i = p_i(\theta) \text{ vs. } H_a : \text{o.w.},$$

one has to estimate the unknown parameter(s) θ then calculate the *estimated expected* $\hat{E} = np_i(\hat{\theta})$.

H_0 will be rejected if

$$\chi^2 = \sum \frac{(O - \hat{E})^2}{\hat{E}} \geq \chi^2_{\alpha,k-1-d},$$

where d is the number of parameters to be estimated ($\dim(\theta)$).

H_0 has d df; H_a has $k - 1$.

Testing for Distributional Models

The sex-ratio data of 72069 six-child families are given below.

Boys	O	$\hat{E}$
0	1096	939.5
1	6233	5982.5
2	15700	15873.1
3	22221	22461.8
4	17332	17879.3
5	7908	7590.2
6	1579	1342.6

The boy ratio is estimated to be $\hat{p} = .5148723$.

Are the boy counts binomial?

$$H_0 : p_i = C_i^6 p^i (1-p)^{6-i}$$

$$H_a : \text{otherwise}$$

H_0 has 1 df; H_a has 6 df.

First calculate the *estimated* expected cell counts

$$\hat{E}_i = nC_i^6 \hat{p}^i (1-\hat{p})^{6-i}.$$

Then calculate

$$\chi^2 = \sum_{i=0}^{6} \frac{(O_i - \hat{E}_i)^2}{\hat{E}_i} = 112.7.$$

Since $P(\chi^2_5 > 112.7) = 0_+$, evidence is overwhelming against a binomial model.

2×2 Table: Homogeneity – I

In a study to evaluate the effectiveness of the drug Timolol in preventing angina attacks, patients were randomly allocated to receive Timolol or placebo for 28 weeks.

	Timolol	Placebo
A-free	44	19
Not A-free	116	128

One has $\hat{p}_1 = \frac{44}{44+116} = .275$, $\hat{p}_2 = \frac{19}{19+128} = .129$, and

$$\sqrt{\frac{.275(.725)}{160} + \frac{.129(.871)}{147}} = .045.$$

A 95% CI for $p_1 - p_2$ is thus

$$(.275 - .129) \pm 1.96(.045),$$

or $(.058, .234)$.

Consider $X_i \sim \text{Bin}(n_i, p_i)$, $i = 1, 2$. For n_1, n_2 large,

$$\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \sim N(0,1),$$

which can be used to construct CI for $p_1 - p_2$.

To test the hypotheses

$$H_0 : p_1 = p_2 \quad \text{vs.} \quad H_a : p_1 \neq p_2,$$

calculate

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}},$$

where $\hat{p} = (X_1 + X_2)/(n_1 + n_2)$, and reject H_0 when $|Z| > z_{\alpha/2}$.

- Plug in H_0 in $\hat{\sigma}_{\hat{p}_1 - \hat{p}_2}$ for test.

2×2 Table: Homogeneity – II

For the angina data, the pooled estimate is $\hat{p} = \frac{44+19}{160+147} = .205$, and

$$\sqrt{.205(.795)(\tfrac{1}{160} + \tfrac{1}{147})} = .046.$$

To test $H_0 : p_1 = p_2$, calculate

$$Z = \frac{.275 - .129}{.046} = 3.159,$$

with p-value $P(|Z| > 3.16) = .0016$.

The estimated expected cell counts under H_0 are

	Timolol	Placebo
A-free	32.83	30.17
Not A-free	127.17	116.83

One calculates

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 9.98 = 3.159^2.$$

One can also use the χ^2-test for $H_0 : p_1 = p_2$. Under H_0, the expected cell counts are

	trt1	trt2
s	$n_1 p$	$n_2 p$
f	$n_1(1-p)$	$n_2(1-p)$

with p estimated by $\hat{p} = \frac{Y_1+Y_2}{n_1+n_2}$. The resulting $\chi^2 = Z^2$.

Write the observed table as

n_{11}	n_{12}	$n_{1\cdot}$
n_{21}	n_{22}	$n_{2\cdot}$
$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot\cdot}$

The estimated expected table consists of $e_{ij} = n_{i\cdot} n_{\cdot j} / n_{\cdot\cdot}$.

2×2 Table: Conditional Probability

To study the relationship between hair color and eye color in a German population, an anthropologist observed a sample of 68000 men.

	Hair		
Eye	Dark	Light	Total
Dark	726	131	857
Light	3129	2814	5943
Total	3855	2945	6800

Simple calculation yields

$$\hat{P}(DE|DH) = \frac{726}{3855} = .1883,$$

$$\hat{P}(DE|LH) = \frac{131}{2945} = .0445,$$

$$\hat{P}(DH|DE) = \frac{726}{857} = .8471.$$

2×2 tables also come up with two binary r.v.'s.

	B	$\bar{\text{B}}$	
A	p_{11}	p_{12}	$p_{1\cdot}$
$\bar{\text{A}}$	p_{21}	p_{22}	$p_{2\cdot}$
	$p_{\cdot 1}$	$p_{\cdot 2}$	1

It is clear that $P(A|B) = p_{11}/p_{\cdot 1}$, $P(B|A) = p_{11}/p_{1\cdot}$, etc.

Estimation of conditional probabilities is straightforward.

$$\hat{P}(A|B) = n_{11}/n_{\cdot 1},$$

$$\hat{P}(A|\bar{B}) = n_{12}/n_{\cdot 2},$$

$$\hat{P}(B|A) = n_{11}/n_{1\cdot}.$$

2×2 Table: Independence

For the hair color and eye color data above,

$$\hat{P}(DE) = \frac{857}{6800} = .1260,$$

$$\hat{P}(DH) = \frac{3855}{6800} = .5669.$$

The estimated expected are

	Hair		
Eye	Dark	Light	Total
Dark	485.8	371.2	857
Light	3369.2	2573.8	5943
Total	3855	2945	6800

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 313.6.$$

Evidence is overwhelming against $H_0 : p_{ij} = p_{i\cdot}p_{\cdot j}$.

For A and B indep., $P(A \cap B) = P(A)P(B)$, or $p_{11} = p_{1\cdot}p_{\cdot 1}$.

To test the hypotheses

$$H_0 : p_{ij} = p_{i\cdot}p_{\cdot j} \quad \text{vs.} \quad H_a : \text{o.w.}$$

estimate $p_{i\cdot}$ by $\hat{p}_{i\cdot} = n_{i\cdot}/n_{\cdot\cdot}$, $p_{\cdot j}$ by $\hat{p}_{\cdot j} = n_{\cdot j}/n_{\cdot\cdot}$, and calculate the estimated expected under H_0,

$$e_{ij} = n_{\cdot\cdot}\hat{p}_{i\cdot}\hat{p}_{\cdot j} = n_{i\cdot}n_{\cdot j}/n_{\cdot\cdot}.$$

- Different problem settings yield the same χ^2.
- For homogeneity, H_0 has 1 df, H_a has 2. For independence, H_0 has 2 df, H_a has 3.

Testing with $r \times c$ Table

Blood types were determined for 1655 ulcer patients and 10000 healthy controls.

	Ulcer	Control	Total
O	911	4578	5489
A	579	4219	4798
B	124	890	1014
AB	41	313	354
Ttl	1655	10000	11655

E's are easily calculated, e.g.,

$$e_{11} = \frac{5489(1655)}{11655} = 779.4.$$

Since $\chi^2 = 49$ and $\chi^2_{.01,3} = 11.34$, reject homogeneity at the 1% level.

A $r \times c$ table can be r outcomes cross c treatments, or the joint distribution of two discrete r.v.'s.

To test for homogeneity or independence, calculate the estimated expected by

$$e_{ij} = n_{i\cdot}n_{\cdot j}/n_{\cdot\cdot}.$$

The test statistic $\chi^2 = \sum \frac{(O-E)^2}{E}$ has $(r-1)(c-1)$ df.

For homogeneity, H_0 has $r-1$ df, H_a has $c(r-1)$. For independence, H_0 has $(r-1)+(c-1)$ df, H_a has $rc-1$.