Chapter 9. INFERENCES CONCERNING PROPORTIONS

April 7, 2011

Example A socialist believes that the percentage of boys in a city is much higher than the girls. He randomly observed 36 children under 5 years old. He denote 1 for boy and 0 for girls and recorded the following data 1, 1, 0, 0, 1, 0, 0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1, 1, 1, 1, 0, 1.

- find the 95% CI for the proportion of boys in the city.
- with significance level 0.05, can he reject the claim by the government that there is no big difference in the gender proportions.

1 Estimation of Proportions

- We are interested in the proportions or percentages of a specified category, or probabilities of a event.
- The information that is usually available for the estimation of a proportion is the number of times, X, that an appropriate event occurs in n trials, occasions, or observations.
- The point estimator of the population proportion itself is usually the sample proportion X/n
- $$E(\frac{X}{n}) = p, \qquad Var(\frac{X}{n}) = \frac{p(1-p)}{n}$$

- Point estimator of p

$$\hat{p} = \frac{X}{n}$$

2 Confidence Interval for p

- When n is large, by the CLT,

$$\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \approx N(0,1), \qquad \text{and} \qquad \frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}} \approx N(0,1)$$

Thus for confidence $100(1-\alpha)\%$, we have

$$P(|\frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}}| < z_{\alpha/2}) \approx 1 - \alpha$$

i.e.

$$P\Big(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\Big) \approx 1 - \alpha$$

Therefore the CI with $100(1-\alpha)\%$ is

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Example If x=36 of n=100 persons interviewed are familiar with the tax incentives for installing certain energy-saving devices, construct a 95% confidence interval for the corresponding true proportion.

Solution: $\hat{p} = x/n = 36/100 = 0.36$. Hence the CI is

$$0.36 \pm z_{\alpha/2}\sqrt{\frac{0.36(1-0.36)}{100}} = [0.266, 0.454]$$

Maximum Estimation error and Sample size

- The error when we use X/n as estimator of p is given by $|X/n - p|$
- Again using the (approximately) normal distribution, we can assert with probability $1 - \alpha$ that the inequality

$$|\hat{p} - p| \leq z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- the Maximum estimation error is

$$E = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- If E is specified in advance, we need n below to achieve the error

$$n = \hat{p}(1-\hat{p})\left(\frac{z_{\alpha/2}}{E}\right)^2$$

Example In a sample survey conducted in a large city, 136 of 400 persons answered yes to the question of whether their city's public transportation is adequate. With 99% confidence, what can we say about the maximum error if x/n=0.34 is used as an estimate of the corresponding true proportion? if the maximum estimation error is confined as 0.05 with the same confidence, what is the appropriate sample size?

- $$E = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 2.575\sqrt{\frac{0.34 * 0.66}{400}} = 0.061$$
- $E = 0.05$,
$$n = \hat{p}(1-\hat{p})\left(\frac{z_{\alpha/2}}{E}\right)^2 = 595.16$$

  596 samples are needed.

3 Hypotheses Concerning One Proportion

- The test of null hypothesis that a proportion equals some specified constant is widely used in sampling inspection, quality control, and reliability verification.

- Null hypothesis $H_0 : p = p_0$

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \approx N(0,1)$$

-

Null hypothesis	Alternative hypothesis	Reject null hypothesis if
$H_0 : p = p_0$	$H_1 : p < p_0$	$z < -z_\alpha$
$H_0 : p = p_0$	$H_1 : p > p_0$	$z > z_\alpha$
$H_0 : p = p_0$	$H_1 : p \neq p_0$	$z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$

Example In a study designed to investigate whether certain detonator used with explosives in coal mining meet the requirement that at least 90% will ignite the explosive when charged, it is found that 174 of 200 detonators function properly. Test the null hypothesis p=0.9 again the alternative p< 0.9 at the 0.05 level of significance.

- Step 1. Testing hypotheses: $H_0 : p = 0.9$ versus $H_1 : p < 0.9$
- Step 2. Level of significance: $\alpha = 0.05$
- Step 3. test statistic

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

Criterion: Reject the null hypothesis if $Z < -1.65$

- Step 4. calculate the z value

$$z = \frac{(174/200 - 0.9)}{\sqrt{0.9 * 0.1/200}} = -1.41$$

- The null hypothesis cannot be rejected
- P-value: $P(Z < -1.41) = 0.079 >$ level of significance 0.01, we also don't reject H_0

4 Statistic for test concerning difference between two proportions

Suppose we have two populations with proportion p_1 and p_2 respectively. We are interested in the relation between p_1 and p_2.

- By the CLT, we have

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

- If $p_1 = p_2$, one can estimate p_1 or p_2 better by the pooled estiamtor

$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

- If $p_1 = p_2$, by the CLT, we have

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}} \approx N(0,1) \tag{4.1}$$

CI for the difference $p_1 - p_2$

with confidence interval $100(1-\alpha)\%$, because

$$P\left(\left| \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \right| < z_{\alpha/2} \right) = 1 - \alpha$$

or

$$P\Big(\hat{p}_1 - \hat{p}_2 - z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} < (p_1 - p_2) < \hat{p}_1 - \hat{p}_2 + z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\Big) = 1 - \alpha$$

i.e. the CI is

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Hypothesis test For testing hypothesis

$$H_0 : p_1 = p_2$$

If H_0 is correct, then

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}} \approx N(0,1)$$

Thus, the rejection region (with significance level α) and p-values are listed below for different alternatives

H_1	rejection region	p-value
$p_1 > p_2$	$z > z_\alpha$	1 - F(z)
$p_1 < p_2$	$z < -z_\alpha$	F(z)
$p_1 \neq p_2$	$\|z\| > z_{\alpha/2}$	2F(-$\|z\|$)

Example A study shows that 16 of 200 tractors produced on one assembly line required extensive adjustments before they could be shipped, while the same was true for 14 of 400 tractors produced on another assembly line.

- Find the large sample 95% confidence interval for $p_1 - p_2$.

- At the 0.01 level of significance, does this support the claim that the second production line does superior work?

SOLUTION

- $\hat{p}_1 = 0.08, \hat{p}_2 = 0.035$

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2}$$

$$= 0.08 - 0.035 \pm 1.96\sqrt{0.08 * 0.92/200 + 0.035 * 0.965/400}$$

$$= [0.003, \quad 0.087]$$

The first assembly line has a rate of extensive adjustment between 3 and 87 out of 1,000, higher than the rate for the second assembly line.

- 1. $H_0 : p_1 = p_2$, v.s. $H_1 : p_1 > p_2$

2. Level of significance: $\alpha = 0.01$

3. Criterion: Reject the null hypothesis if $Z > 2.33$, where Z is given by above in (4.1).

4. Calculations:

$$\hat{p} = \frac{X_1 + X_2}{200 + 400} = 0.05$$

with $n_1 = 200, n_2 = 400$. Substituting into the Z statistic, we have

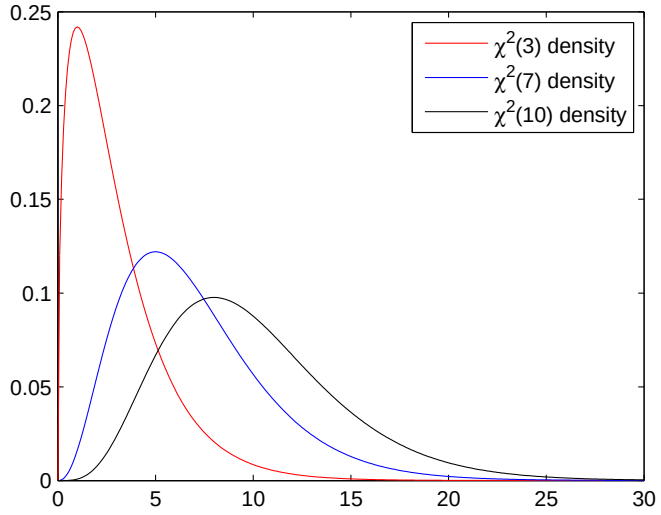
$$z = 2.38$$

5. Decision: Since $z = 2.38$ exceeds 2.33, the null hypothesis must be rejected; we conclude that the true proportion of tractors requiring extensive adjustments is greater for first assembly line than for the second. (P-value = 0.0087)

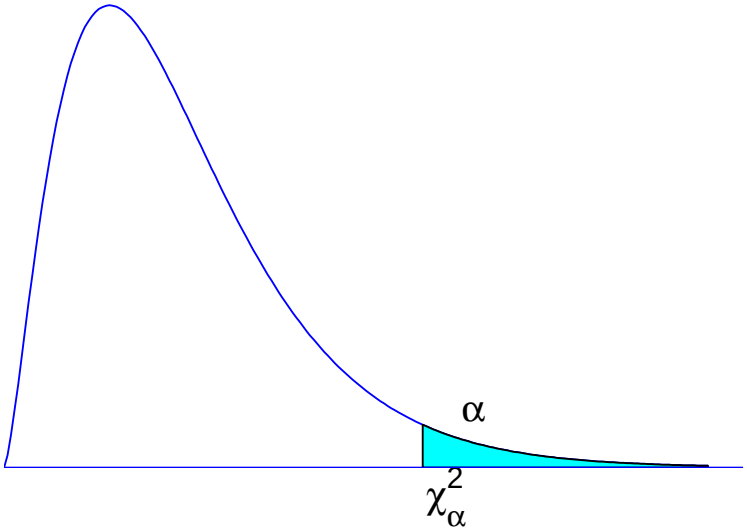
5 χ^2 distribution (for more discussion, see Chapter 6)

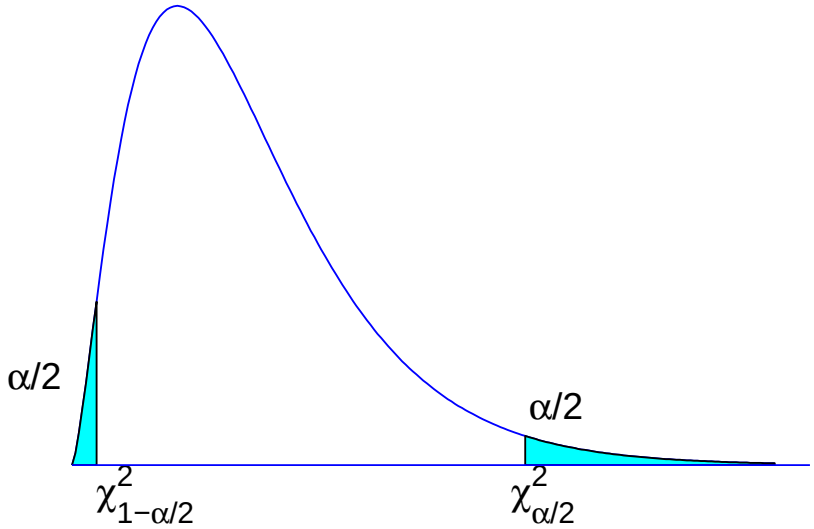
A random variable follows $\chi^2(k)$ distribution if it has the following p.d.f.

1. the random is nonnegative
2. the density looks as follows, depending the number of k, called the degree of freedom



3. the following critical values will be used later





6 Hypotheses Concerning Several Proportions

- **Example** We compare the consumer response to two different products, when we decide whether the proportion of defectives of a given process remains constant from day to day.
- Generally, we have k populations, each has proportion $p_i, i = 1, 2, ..., k$.
- Thus, we are interested in testing

$$H_0 : p_1 = p_2 = ... = p_k = p$$

versus

$$H_1 : p_1, p_2, ..., p_k \text{ are not all the same}$$

Large Sample test to compare one proportion

- We require **independent random samples** of size $n_1, n_2, ..., n_k$ from each population respectively. If the corresponding number of successes are $X_1, X_2, ..., X_k$, the test we should use is based on the fact that

  1. Large samples the sampling distribution of

  $$Z_i = \frac{X_i/n_i - p_i}{\sqrt{p_i(1-p_i)/n_i}}$$

  is approximately the standard normal distribution

  2. (proof is not required)

  $$\chi^2 = \sum_{i=1}^{k} \frac{(X_i - n_i p_i)^2}{n_i p_i (1-p_i)}$$

is a value of random variable having approximately the chi-square distribution with k degrees of freedom.

3. In practice, we substitute for the p_i, which under the null hypothesis are all equal, the pooled estimate

$$\hat{p} = \frac{X_1 + X_2 + ... + X_k}{n_1 + n_2 + ... + n_k}$$

4. The null hypothesis should be rejected if the difference between the X_i and $n_i\hat{p}$ are large, the critical region is

$$\chi^2 > \chi^2_\alpha$$

where the number of degrees of freedom is k-1.

Another approach (more general)

To include the case of compare 2 or more proportions,

- consider the following table (called contingency table)

	Sample 1 $(j = 1)$	...	Sample k $(j = k)$	Total
Sucesses $(i = 1)$	x_1	...	x_k	x
Failures $(i = 2)$	$n_1 - x_1$	...	$n_k - x_k$	$n - x$
total	n_1	...	n_k	n

- Define the observed cell frequency

$$o_{ij} : i = 1, 2, \quad j = 1, ..., k$$

- with H_0 being true, the expected number of successes and failures for the jth sample are estimated by

$$e_{1j} = n_j \hat{p} \quad \text{and} \quad e_{2j} = n_j(1 - \hat{p})$$

Thus

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

will follows a χ^2 distribution with degree of freedom $k - 1$

Example

Samples of three kinds of materials, subjected to extreme temperature changes. produced the results shown in the following table:

	Material A	Material B	Material C	Total
Crumbled	41	27	22	90
Remained intact	79	53	78	210
total	120	80	100	300

Use the 0.05 level of significance to test whether under the stated conditions, the probability of crumbling is the same for the three kinds of materials.

- $H_0 : p_1 = p_2 = p_3, \quad H_1 : p_1, p_2, p_3$ are not all equal
- Level of significance: $\alpha = 0.05$
- Criterion : Reject the null hypothesis if $\chi^2 > 5.991$, the value of $\chi^2_{0.05}$ for 3 - 1 = 2 degrees of freedom.
- Calculations: The expected frequencies for the first two cells o f the first row are

$$e_{11} = 120 * \frac{90}{300} = 36, \quad e_{12} = 80 * \frac{90}{300} = 24 \quad e_{13} = 100 * \frac{90}{300} = 30$$

and

$$e_{21} = 120 * \frac{210}{300} = 84, \quad e_{22} = 80 * \frac{210}{300} = 56 \quad e_{23} = 100 * \frac{210}{300} = 70$$

Thus the value

$$\begin{aligned}\chi^2 &= \frac{(41-36)^2}{36} + \frac{(27-24)^2}{24} + \frac{(41-30)^2}{30} \\ &\quad + \frac{(79-84)^2}{84} + \frac{(53-56)^2}{56} + \frac{(78-70)^2}{70} \\ &= 4.575\end{aligned}$$

- Decision: Since $\chi^2 = 4.575$ does not exceed 5.991. The null hypothesis cannot be rejected. in other words, the data do not refute the hypothesis that, under the stated conditions, the probability of crumbling is the same for the three kinds of material.