Matematisk Statistik och Disktret Matematik, MVE051/MSG810, VT19

Föreläsning 10

Nancy Abdallah

Chalmers - Göteborgs Universitet

Estimating proportions

Example

Suppose we want to estimate the proportion of people who own tablets in a certain city. 250 randomly selected people are surveyed, 98 of them reported owning tablets. An estimate for the population proportion is given by $\hat{p} = \frac{98}{250} = 0.392$.

In general we want to study a particular trait in a population. We ask about the proportion of the population with this trait, or, more precisely, an estimate of this proportion.

- We choose a random sample $X_1, \ldots, X_n$ from the population.
- The variables are independent and can be defined as follows:

$$X_i = \begin{cases} 1 & \text{if the } i\text{th member of the sample has the trait} \\ 0 & \text{otherwise} \end{cases}$$

- the point estimator for p, the population proportion, is

$$\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n}$$

- $\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n}$ is an unbiased estimator for p. To prove that, we notice that $p(X_i = 1) = p$. Then

$$E[X_i] = 1 \cdot p + 0 \cdot (1 - p)$$

$$\Rightarrow E[\hat{p}] = \frac{\sum_{i=1}^{n} E[X_i]}{n} = \frac{np}{n} = p$$

- Variance of $\hat{p}$:

$$Var(X_i) = E[X_i^2] - E[X_i]^2 = p - p^2 = p(1 - p)$$

$$\Rightarrow V[\hat{p}] = \frac{\sum V[X_i]}{n^2} = \frac{np(1 - p)}{n^2} = \frac{p(1 - p)}{n}$$

  Hence, when n is large, the variance is too small.

This makes $\hat{p}$ a good estimator.

C.I. on the proportion

- When we take n large enough, by the central limit theorem, $\hat{p}$ is approximately normally distributed with mean p and variance $p(1-p)/n$.
- A $100(1-\alpha)\%$ confidence interval is defined by

$$(\hat{p} - z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}, \hat{p} + z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n})$$

where $P(x \leq z_{\alpha/2}) = 1 - \alpha/2$.

Example

A 95% C.I. on the proportion of people who own a tablet (see example p.1) is given by

$$\left(0.392 - 1.95\sqrt{\frac{0.392(0.608)}{250}}, 0.392 + 1.64\sqrt{\frac{0.392(0.608)}{250}}\right)$$

$$= (0.341, 0.443)$$

Sample size for estimating p

- The points in the confidence interval $(\hat{p} - z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}, \hat{p} + z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n})$ lie within a distance of
$$d = z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} \tag{1}$$
from the mean.
- Using Equation 1 we get,
$$n = \frac{z^2_{\alpha/2}\hat{p}(1-\hat{p})}{d^2}$$
which is the sample size for estimating p with prior estimate available.

Sample size for estimating *p*

- If no prior estimate for *p* is available, the sample size needed for estimating *p* is given by

$$n = \frac{z_{\alpha/2}^2}{4d^2}$$

Example

A mobile phone company wants to determine the current percentage of customers ages 50+ who use text messaging on their cell phones. How many customers ages 50+ should the company survey in order to be 90 percent confident that the estimated (sample) proportion is within 3 percentage points of the true population proportion of customers ages 50+ who use text messaging on their cell phones assuming that

1. $\hat{p} = 0.4$ is a prior estimate for p.
2. no prior estimate is given

Solution $d = 0.03$, $z_{0.05} = 1.645$.

1. $n = \frac{z^2_{\alpha/2}\hat{p}(1-\hat{p})}{d^2} = \frac{1.645^2(0.4)(0.6)}{0.03^2} \approx 721$
2. $n = \frac{1.645^2}{4(0.03^2)} \approx 752.$

Hypothesis testing on proportion

- In the same way as for hypothesis testing on the mean, we can do a hypothesis testing on the population proportion.
- The test statistic is

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

  where p_0 is the value of p used in the hypotheses.
- If n is large enough, the test statistic follows an approximate normal distribution.
- n is consider large enough if

$$p_0 \leq 0.5 \text{ and } np_0 > 5$$

  or

$$p_0 > 0.5 \text{ and } n(1 - p_0) > 5$$

Example

Newborn babies are more likely to be boys than girls. A random sample found 13 173 boys were born among 25 468 newborn children. The sample proportion of boys was 0.5172. Is this sample evidence that the birth of boys is more common than the birth of girls in the entire population? Let $\alpha = 0.05$.

Solution:

$$H_0 : p = 0.5 \text{ and } H_1 : p > 0.5$$

Since n is large, $z = \frac{\hat{p}-0.5}{\sqrt{0.5(0.5)/25468}}$ is approximately normally distributed.

The critical value is $z_{0.05} = 1.645$ and $z = \frac{0.5172-0.5}{\sqrt{0.5(0.5)/25468}} = 5.49$ which is in the rejection region. Therefore H_0 is rejected and hence the sample gives evidence that the proportion of boys is higher than that of girls.

Comparing two proportions

Suppose we have two populations and we are interested in a certain trait. The proportion of the members having the trait in either population is unknown. We want to estimate these proportions and the difference between them in order to compare the two populations.

Example

We are interested in comparing the proportion of researchers who use a certain computer program in their research in two different fields: pure mathematics and probability and statistics.
Populations: Researchers in the pure math field and researchers in the probability and statistics field .
Trait of interest: Usage of computer programs.

Point estimator and C.I for the difference between two proportions

Suppose that p_1 is the true proportion of population 1 and p_2 is that of population 2.

- From each population we take a random sample such that the samples are independent from each other.
- For each sample we compute the point estimate: $\hat{p}_1$ and $\hat{p}_2$.
- A point estimator for $p_1 - p_2$ is $\hat{p}_1 - \hat{p}_2$.
- For large samples, $\hat{p}_1 - \hat{p}_2$ is approximately normal with mean $p_1 - p_2$ and variance $p_1(1-p_1)/n_1 + p_2(1-p_2)/p_2$, where and n_1 and n_2 are the sample sizes from population 1 and 2 respectively
- A $100(1-\alpha)\%$ C.I. on $p_1 - p_2$ is given by

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

Example

We take a sample of size 375 from population 1 and 375 from population 2. The number of researchers that use a computer program we get from population 1 is 195 and that of researchers from population 2 is 232. Then $\hat{p}_1 = \frac{195}{375} = 0.52$ and $\hat{p}_2 = \frac{232}{375} = 0.619$.
A point estimate for the difference $p_1 - p_2$ is 0.52-0.619=-0.099.
The standard deviation is

$$\sqrt{0.52(0.48)/375 + 0.619(0.381)/375} = 0.036$$

Example

A 95% confidence interval for $p_1 - p_2$ is

$$(0.52 - 0.619 - 1.96(0.036), 0.52 - 0.619 + 1.96(0.036))$$

$$(-0.17, -0.028)$$

Since the interval does not contain 0 and is negative–valued, we can say with 95% level of confidence that the proportion of researchers from population 2 is higher than that of population 1.