Lecture 11: Estimating proportions

MVE055 / MSG810 Mathematical statistics and discrete mathematics)

Moritz Schauer

Last updated September 30, 2020, 2020

GU & Chalmers University of Technology

Estimating proportions

Estimating proportions

Example

Suppose we want to estimate the proportion p 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 $\hat{p} = \frac{38}{250} = 0.392.$

In general we want to study a particular trait in a population too large to sample completely. We ask about the proportion of the population with this trait.

Estimating a proportion

- We choose a random sample $X_1, ..., X_n$ from the population.
- $$X_i = \begin{cases} 1 & i\text{th member of the sample has the trait} \\ 0 & \text{otherwise} \end{cases}$$
- The point estimator is based on the

$$\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n} \quad \text{(proportion in the sample)} \quad .$$

Bernouli random variables

Why do we write $\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n}$ as sum of random variables.

$\mathrm{P}(X_i = 1) = p$, $\mathrm{P}(X_i = 0) = 1 - p$. X_i are Bernoulli random variables with parameter p!

We know a lot about them. E.g.

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

$\hat{p}$ is the some of Bernoulli random variables, hence $\mathrm{Bin}(n, p)$ distributed. So ...

Unbiasedness

Unbiasedness

The expectation of $\hat{p}$:

$$\mathsf{E}[\hat{p}] = \frac{1}{n}\sum_{i=1}^{n}\mathsf{E}[X_i] = \frac{1}{n}\underbrace{(p + p + \cdots + p)}_{n \text{ times}} = p$$

$$\mathsf{E}[\hat{p}] = p$$

$\hat{p}$ is an unbiased estimator for the proportion p.

Variance

The variance of $\hat{p}$ tells us how good as estimator $\hat{p}$ is.

$\operatorname{Var}(X_i) = \mathsf{E}\left[X_i^2\right] - \mathsf{E}\left[X_i\right]^2 = p - p^2 = p(1-p)$

$\Rightarrow \operatorname{Var}(\hat{p}) = \frac{\sum \operatorname{Var}(X_i)}{n^2} = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$

Standard error

The variance of $\hat{p}$:

$$\operatorname{Var}(\hat{p}) = \frac{p(1-p)}{n}$$

The standard error is

$$\mathrm{SE} = \sqrt{\operatorname{Var}(\hat{p})} \approx \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}$$

How many more observations do I need to reduce the standard error by a factor 2? 4 times as much

Example (ctd.)

Recall $\hat{p} = \frac{38}{250} = 0.392$.

The standard error the estimated proportion of people who own a tablet is

$$\mathrm{SE} = \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} = \frac{\sqrt{0.392(1-0.392)}}{\sqrt{250}} = \frac{0.392(0.608)}{250}$$

Confidence interval on $\hat{p}$.

Normal approximation When we take n large enough, by the central limit theorem, $\hat{p}$ is approximately normally distributed with mean p and variance $p(1-p)/n$.

Confidence interval

A $100(1-\alpha)\%$ confidence interval is defined by

$$\left(\hat{p} - z_{\alpha/2}\mathrm{SE}, \hat{p} + z_{\alpha/2}\mathrm{SE}\right)$$

where $\mathrm{SE} = \sqrt{\hat{p}(1-\hat{p})/n}$ and $\mathrm{P}\left(-z_{\alpha/2} \leq x \leq z_{\alpha/2}\right) = 1 - \alpha/2$.

E.g. for a $95\,\%$ CI $z_{\alpha/2} = 1.96$.

Example (ctd.)

A $95\,\%$ C.I. on the proportion of people who own a tablet is given by $\left(\hat{p} - z_{\alpha/2}\mathrm{SE}, \hat{p} + z_{\alpha/2}\mathrm{SE}\right)$ where $\hat{p} = \frac{38}{250}$, $z_{\alpha/2} = 1.96$, $SE = \frac{0.392(0.608)}{250}$.

$$\left(0.392 - 1.96\sqrt{\frac{0.392(0.608)}{250}}, 0.392 + 1.96\sqrt{\frac{0.392(0.608)}{250}}\right)$$

$= (0.3315, 0.4525)$.

"We are 95% confident that proportion of people owning a tablet is somewhere in the interval $(0.3315, 0.4525)$."

Hypothesis test for hypothesis about proportion

We can test hypotheses about the a population proportion:

$$H_0 : p = p_0 \quad \text{and} \quad H_1 : p \underset{<}{\overset{\neq}{>}} p_0$$

Our test statistic is the z-value

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

where p_0 is the null value, the value of p used in the null hypotheses.

The corresponding r.v. Z is approximately standard normal distributed for large n.

Minimum sample size

n is considered large enough if $np_0 > 5$ and $n(1 - p_0) > 5$ (both).

Example

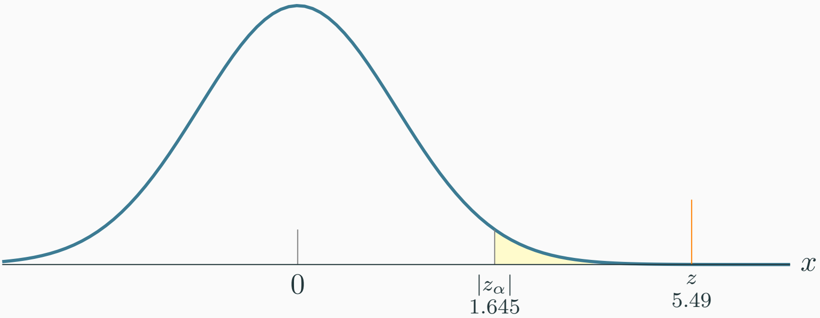
Newborn babies are more likely to be boys than girls. A random sample found 13 173 boys were born among 25 468 newborn children. The sample proportion of boys was 0.5172. Is this sample evidence that the birth of boys is more common than the birth of girls in the entire population? Let $\alpha = 0.05$.

Test

$$H_0\colon p = 0.5 \quad \text{and} \quad H_1\colon p > 0.5.$$

at significance level $\alpha = 0.05$.

Since n is large, $z = \frac{\hat{p}-0.5}{\sqrt{0.5(0.5)/25468}}$ is approximately normally distributed. The critical point is $z_{0.95} = 1.645$ 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.



Rejection region for $\alpha = 0.05$ (on the x-axis below the yellow area).

Comparing two proportions

Suppose we have two populations and we want to compare the proportions in the populations that have a certain trait. Denote the unknown proportions p_1 and p_2.

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 the computer program.

Point estimator and SE 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 of sizes n_1, n_2 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 $p_1 \left(1 - p_1\right) / n_1 + p_2 \left(1 - p_2\right) / p_2$ where and n_1 and n_2 are the sample sizes from population 1 and 2 respectively.

Confidence interval

Confidence interval

A $100(1-\alpha)\%$ C.I. on $p_1 - p_2$ is given by
$\left(\hat{p} - z_{\alpha/2}\mathrm{SE}, \hat{p} + z_{\alpha/2}\mathrm{SE}\right) =$

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

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 (ctd.)

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.