Inferences on Proportions

Estimating Proportions

One of the (mathematically) simplest population parameters of general interest is the ***proportion*** of members of a population with some trait. Every member of the population is characterized as either having or not having this trait. We describe this mathematically by defining the random variable

$$X = \begin{cases} 1 & \text{has trait,} \\ 0 & \text{does not have trait.} \end{cases}$$

The proportion of the members of the population having the trait is

$$p = \frac{\#\text{ members wih trait}}{\text{population size}} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

where N is the population size and x_i is the value of the variable X for the ith member of the population. Hence the proportion is equal to the mean of X.

Estimating Proportions

It follows that if we take a random sample $X_1, \ldots, X_n$ of X, the sample mean

$$\widehat{p} = \overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

is an (unbiased) estimator for p.

The random variable X follows a Bernoulli distribution with expectation $\mathrm{E}[X] = p$ and variance $\mathrm{Var}\, X = p(1-p)$. However, it is often more convenient to use a normal distribution for confidence intervals and hypothesis tests.

By the central limit theorem, $\widehat{p}$ is approximately normally distributed with mean p and variance $p(1-p)/n$. Hence,

$$\frac{\widehat{p} - p}{\sqrt{p(1-p)/n}}$$

is approximately standard-normally distributed.

Estimating Proportions

It follows immediately that the following is a $100(1-\alpha)\%$ confidence interval for p:

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

But the interval depends on the unknown parameter p, which we are actually trying to estimate! One solution to the problem is to replace p by $\widehat{p}$, i.e., to write

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

But then the number $z_{\alpha/2}$ is no longer accurate (when we replaced σ by S to obtain a confidence interval for the mean, we had to switch from $z_{\alpha/2}$ to $t_{\alpha/2}$).

However, we are approximating the binomial distribution in any case - we might argue that if the sample size n is large enough to allow the central limit theorem to hold, then the difference between $z_{\alpha/2}$ and a corrected value will be negligible. This is not a perfect solution, but a detailed discussion would lead to far here.

Estimating Proportions

19.1. Example. In 2017, the Shanghai Academy of Social Sciences Institute of Sociology conducted a survey among residents of Shanghai to ask their opinion about the municipal proposal to limit the numbers of residents to 25 million by 2020 and keep that number stable until 2040.

Among the 2079 residents surveyed, 48.5% indicated that this measure would benefit Shanghai's development. Assuming that those questioned constitute a random sample of Shanghai residents, a 99% confidence interval for the proportion of residents with this opinion is given by

$$p = 0.485 \pm 2.575\sqrt{0.485 \cdot 0.515/2079} = 0.485 \pm 0.028$$

Literature: `https://archive.shine.cn/metro/society/Pros-and-cons-of-limiting-citys-population/shdaily.shtml`

Choosing the Sample Size

As a practical matter, we are often able to choose (perhaps within constraints) the sample size. We may want to be able to claim that "with $xx\%$ probability, $\widehat{p}$ differs from p by at most d."

Given a $100(1-\alpha)\%$ confidence interval $p = \widehat{p} \pm z_{\alpha/2}\sqrt{\widehat{p}(1-\widehat{p})/n}$, we know with $100(1-\alpha)\%$ confidence that

$$d = z_{\alpha/2}\sqrt{\widehat{p}(1-\widehat{p})/n}.$$

Given d, this means that we should choose

$$n = \frac{z_{\alpha/2}^2 \widehat{p}(1-\widehat{p})}{d^2}$$

to ensure that $|p - \widehat{p}| < d$ with $100(1-\alpha)\%$ confidence. However, this formula requires us to have an estimate $\widehat{p}$ of p beforehand.

Choosing the Sample Size

If no estimate for p is available, we can at least use that $x(1-x) < 1/4$ for all $x \in \mathbb{R}$ to deduce that

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

will ensure $|p - \widehat{p}| < d$ with $100(1-\alpha)\%$ confidence.

19.2. Example. A new method of pre-coating fittings used in oil, brake and other fluid systems in heavy-duty trucks is being studied. How large a sample is needed to estimate the proportion of fittings that leak to within 0.02 with 90% confidence?

Since no prior estimate is available, we take

$$n = \frac{z_{0.05}^2}{4d^2} = \frac{1.645^2}{4 \cdot 0.02^2} = 1692.$$

Hypothesis Testing

19.3. Test for Proportion. Let $X_1, \ldots, X_n$ be a random sample of size n from a Bernoulli distribution with parameter p and let $\widehat{p} = \overline{X}$ denote the sample mean. Then any test based on the statistic

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

is called a ***large-sample test for proportion***.

We reject at significance level α

- $H_0 : p = p_0$ if $|Z| > z_{\alpha/2}$,
- $H_0 : p \leq p_0$ if $Z > z_\alpha$,
- $H_0 : p \geq p_0$ if $Z < -z_\alpha$.

Comparing Two Proportions

Two Populations:

- $X^{(1)} \sim \text{Bernoulli}(p_1)$,
- $X^{(2)} \sim \text{Bernoulli}(p_1)$.

Goal: make inferences on $p_1 - p_2$.

Suppose a random sample of size n_1 from population 1 and another random sample of size n_2 from population 2 are given.

An unbiased estimator for $p_1 - p_2$ is

$$\widehat{p_1 - p_2} := \widehat{p}_1 - \widehat{p}_2 = \overline{X}^{(1)} - \overline{X}^{(2)},$$

where $\overline{X}^{(1)}$ and $\overline{X}^{(2)}$ are the sample means of the respective random samples.

A Confidence Interval

We have the approximate distributions

$$\overline{X}^{(1)} \sim N\left(p_1, \frac{p_1(1-p_1)}{n_1}\right), \qquad \overline{X}^{(2)} \sim N\left(p_2, \frac{p_2(1-p_2)}{n_2}\right)$$

We can infer that for large samples

$$\widehat{p_1 - p_2} \sim N\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$$

This allows us to deduce the following $100(1-\alpha)\%$ confidence interval for $p_1 - p_2$:

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

which is valid for large sample sizes.

Comparing Two Proportions

19.4. Test for Comparing Two Proportions. Suppose two samples of (large) sizes n_1 and n_2 from two Bernoulli distributions with parameters p_1 and p_2 are given. Denote by $\widehat{p}_1$ and $\widehat{p}_2$ the means of the two samples.

Let $(p_1 - p_2)_0$ be a null value for the difference $p_1 - p_2$. Then the test based on the statistic

$$Z = \frac{\widehat{p}_1 - \widehat{p}_2 - (p_1 - p_2)_0}{\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}}$$

is called a ***large-sample test for differences in proportions***.

We reject at significance level α

- $H_0 : p_1 - p_2 = (p_1 - p_2)_0$ if $|Z| > z_{\alpha/2}$,
- $H_0 : p_1 - p_2 \leq (p_1 - p_2)_0$ if $Z > z_\alpha$,
- $H_0 : p_1 - p_2 \geq (p_1 - p_2)_0$ if $Z < -z_\alpha$.

Pooled Estimator for the Proportion

Most commonly we test against the null value $(p_1 - p_2)_0 = 0$, i.e.,

$$H_0 : p_1 = p_2.$$

If H_0 is true, the common proportion is

$$p = p_1 = p_2.$$

Then both $\widehat{p}_1$ and $\widehat{p}_2$ are estimators for p.

It turns out that the best course of action is to take the weighted average: we define the ***pooled estimator for the proportion***,

$$\widehat{p} := \frac{n_1\widehat{p}_1 + n_2\widehat{p}_2}{n_1 + n_2}. \tag{19.1}$$

Pooled Test for Equality of Proportions

19.5. Pooled Test for Equality of Proportions. Suppose two samples of (large) sizes n_1 and n_2 from two Bernoulli distributions with parameters p_1 and p_2 are given. Denote by $\widehat{p}_1$ and $\widehat{p}_2$ the means of the two samples. Let $\widehat{p}$ be the pooled estimator for the proportion. Then the test based on the statistic

$$Z = \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

is called a ***pooled large-sample test for equality of proportions***.

We reject at significance level α

- $H_0 : p_1 = p_2$ if $|Z| > z_{\alpha/2}$,
- $H_0 : p_1 \leq p_2$ if $Z > z_\alpha$,
- $H_0 : p_1 \geq p_2$ if $Z < -z_\alpha$.

Pooled Proportions

19.6. Example. Many consumers think that automobiles built on Mondays are more likely to have serious defects than those built on any other day of the week. To support this theory, a random sample of 100 cars built on Monday is selected and inspected. Of these, eight are found to have serious defects. A random sample of 200 cars produced on other days reveals 12 with serious defects. Do these data support the stated contention?

We test

$$H_0 : p_1 \leq p_2.$$

where p_1 denotes the proportion of cars with serious defects produced on Mondays.

Estimates for p_1 and p_2 are

$$\widehat{p}_1 = 8/100 = 0.08, \qquad \widehat{p}_2 = 12/200 = 0.06.$$

Pooled Proportions

The pooled estimate for the common population proportion is

$$\widehat{p} = \frac{100 \cdot 0.08 + 200 \cdot 0.06}{100 + 200} = 20/300 = 0.066.$$

The observed value of the test statistic is

$$\frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.08 - 0.06}{\sqrt{0.066 \cdot 0.934\left(\frac{1}{100} + \frac{1}{200}\right)}} = 0.658.$$

From the standard normal table, we see that the probability of observing this large or a larger value is 0.2546, so there is no evidence that H_0 might be false.