Analysis of Proportions

Valeria Vitelli
Oslo Centre for Biostatistics and Epidemiology
Department of Biostatistics, UiO
valeria.vitelli@medisin.uio.no

MF9130E – Introductory Course in Statistics
11.04.2024

Proportions and the binomial distribution

The binomial distribution function

- The formula for a **binomial probability**, or, the probability of getting exactly d events in a sample of n individuals is

$$P(d \text{ events}) = \binom{n}{d} \pi^d (1 - \pi)^{n-d},$$

  where $d! = 1 \times 2 \times \ldots \times d$, and π is the population probability of the event of interest (D)
- We wish to estimate π using p
- How much do does p differ from the true, unknown π? What is the uncertainty?

Standard error of a proportion

- The **standard error** of the proportion of D's in a sample is:

$$\text{s.e.} = \sqrt{\frac{\pi(1-\pi)}{n}},$$

where n is the sample size. It measures how closely the sample proportion estimates the population proportion

- The standard error is estimated by:

$$\widehat{\text{s.e.}} = \sqrt{\frac{p(1-p)}{n}},$$

with π replaced by p

The normal approximation to the binomial distribution

- When the sample size, n, increases, the binomial distribution can be approximated by a **normal distribution** with the same mean and standard error as for the binomial distribution
- This is useful for:
  - calculating **confidence intervals**
  - carrying out **hypothesis tests**

- **Rule of thumb**: The approximation is valid when both $n \times \pi$ and $n \times (1 - \pi)$ is greater than or equal to 10

Confidence interval for a proportion

- Given that the normal approximation to the binomial distribution is sufficiently good, the **confidence interval for the population proportion**, π, is

$$\mathrm{CI} = \left(p - z' \times \sqrt{\frac{p(1-p)}{n}}, p + z' \times \sqrt{\frac{p(1-p)}{n}}\right),$$

  where z' is the appropriate percentage point of the standard normal distribution (typically 1.96), n is the sample size, and p is the sample proportion

Example: 15.3 in Kirkwood & Sterne

In September 2001 a survey of **smoking habits** was conducted in a sample of 1000 teenagers aged 15-16, selected at random from all 15-16 year-olds living in Birmingham, UK. A total of 123 reported that they were current smokers

Thus the **proportion** of current smokers is estimated by

$$p = \frac{123}{1000} = 0.123 = 12.3\%$$

The **standard error** of p is estimated by:

$$\widehat{\text{s.e.}} = \sqrt{\frac{0.123 \times (1 - 0.123)}{1000}} = 0.0104$$

Thus the 95% **confidence interval** for the population probability is:

$$\begin{aligned} 95\%\ \mathrm{CI} &= (0.123 - 1.96 \times 0.0104, 0.123 + 1.96 \times 0.0104) \\ &= (0.103, 0.143) \end{aligned}$$

This means that with 95% confidence, in September 2001 the proportion of 15-16 year-olds living in Birmingham who smoked was between 10.3% and 14.3%

Testing a hypothesis about one proportion

Hypothesis testing

- To test the **null hypothesis that the population proportion equals a particular value**, π_0:

$$\mathrm{H}_0 : \pi = \pi_0, \mathrm{H}_a : \pi \neq \pi_0,$$

  we perform a *z*-test using the approximating normal distribution

z-test

- Providing that both $n \times \pi_0$ and $n \times (1 - \pi_0)$ are greater than or equal to 10, the **test statistic**

$$z = \frac{p - \pi_0}{\sqrt{\pi_0(1 - \pi_0)/n}}$$

  is **standard normally distributed**
- From the test statistic we derive a corresponding P-**value**, which is the probability that $\pi = \pi_0$ (or something more extreme)

Example: 15.3 in Kirkwood & Sterne

In 1998 the UK Government announced a target of **reducing smoking among children** from the national average of 13% to 9% or less by the year 2010, with a fall to 11% by the year 2005. Is there evidence that the proportion of 15-16 year-old smokers in Birmingham at the time of our survey in 2001 was below the national average of 13% at the time the target was set?

To answer the question, we carry out a statistical test where the **null hypothesis** states that the population proportion is equal to 0.13 (13%), while the alternative states that the population proportion is less than 0.13 (13%):

$$\mathrm{H}_0: \quad \pi = 0.13 \quad \text{vs} \quad \mathrm{H}_1: \quad \pi < 0.13$$

The **standard error** of the sample proportion, *p*, under the null hypothesis is:

$$\sqrt{\frac{0.13 \times (1 - 0.13)}{1000}} = 0.106$$

The observed value of the **test statistic** is therefore:

$$z = \frac{0.123 - 0.13}{0.106} = -0.658$$

with a corresponding (one-sided) ***P*-value** equal to 0.255. This means that there is no evidence that the proportion of teenage smokers in Birmingham in September 2001 was lower than the national 1998 levels