Statistics

Week 8: Inference for Proportions (Chapter 9)

ESD, SUTD

Term 5, 2017

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Established in collaboration with MIT

Due dates

Homework 2 is due at 1pm, Tuesday 14 March.

If you chose **Option 2** for the project, then your 1 page midterm report is **due** this Friday (by 11:59pm).

Please email the report to me. It should include: an outline of the problem you are investigating, what you have done so far (e. g. any data collected), and what you plan to do.

Guest lectures

Professor Roy Welsch (MIT) is visiting us for 4 weeks and will be giving 4 **guest lectures**, held during normal lecture or recitation times.

The material covered in the guest lectures is examinable.

Exam

You will get to see your midterm exam during this Thursday's recitation.

Some of you wrote '$\mathsf{E}(\sqrt{X}) = \sqrt{\mathsf{E}(X)}$'. Is this equality true? It is important to have some intuition for these things.

Outline

1. Inference for a proportion

2. Chi-squared test
   - Multinomial

Proportions and percentages

Often we are interested in proportions or percentages: the % of people who have a certain opinion or belong to a certain group, the % of successes, the % of defective products, . . .

Question

A survey is conducted to estimate the proportion of people who favour a particular political party. The proportion is to be estimated within a *margin of error* of 3 percentage points, with 95% *confidence*. What sample size should be planned, if the population consists of:

- everyone in this room ($N \approx 30$)?
- everyone in Singapore ($N \approx 5.5$ million)?
- everyone in USA ($N \approx 320$ million)?

Set up

Set up: proportion p of a population has a certain attribute, and we wish to estimate p. We take a random sample, $X_1, X_2, \ldots, X_n$, from the population; they can be treated as iid *Bernoulli* random variables.

An unbiased estimator for p is

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

The CLT (or the normal approximation for binomial) tells us that $\hat{p}$ is approximately $N(p, p(1-p)/n)$ for *large* n.

Therefore,

$$P\left(\hat{p} - z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} \le p \le \hat{p} + z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right) \approx 1 - \alpha.$$

When n is large, we can estimate $p(1-p)$ by $\hat{p}(1-\hat{p})$.

Confidence interval

Therefore the $(1-\alpha)$-level, two-sided confidence interval for p is

$$\left[\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

Exercise: write down the one-sided CI's.

Remark: it is possible to solve the inequalities on the last slide for p, using the quadratic formula, and obtain a more accurate CI without resorting to the estimate. However, often in practice, the two CI's give very similar results. See textbook Section 9.1.

Sample size

We call $E := z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ the **margin of error**. Then, the minimum sample size for a given margin of error is

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

In practice, we may use the conservative value

$$n = \left(\frac{z_{\alpha/2}}{2E}\right)^2,$$

rounded up to the next integer. (Why?)

Exercise: answer the questions on slide 6. Some of the answers may be counter-intuitive, so discuss with your neighbours, and defend your answers if needed.

Hypothesis testing

In inference for a proportion, we usually perform hypothesis testing by computing the p-value.

Unlike in inference for a single sample, this is *not* quite equivalent to constructing a confidence interval.

Example

A basketball player has had a long time average of making 70% of attempted free throws. In the current season he makes 297 out of 396 attempted free throws. Has his free throw average actually improved? Use $\alpha = 0.05$.

Hypothesis testing (continued)

Solution: let $p_0 = 0.7$. We have $H_0 : p = p_0, \;\; H_1 : p > p_0$.

Assume H_0 is true, then $\dfrac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} = \dfrac{0.75 - 0.7}{\sqrt{0.7 \times 0.3/396}}$ is approximately an observation from a standard normal distribution.

The p-value $= 0.015 < \alpha$, so we reject H_0.

Notice how in computing the p-value, we have used p_0. If we constructed a CI (not the preferred method for hypothesis testing here), then we would have used $\hat{p}$.

Incidentally, the CI is $[0.714, \, 1)$.

Polls – things to watch out for

The media often use polls with small sample sizes, and sensationalize the random fluctuations. Sometimes, the polls are carried out or reported in a way to advance someone's agenda or to sway public opinion.

Internet polling can be inaccurate, especially when the participants are not representative of the population, or if they do not want to disclose their true opinion for whatever reason.

A recent example is the 2016 US presidential election. One poll expert promised to 'eat a bug' if Trump won more than 240 electoral votes (he won 304).

Watch it here

More remarks

There are formulas for computing the power of the test.

There are also ways to perform inferences for comparing two proportions (some of it can also be done via the chi-squared test if the sample size is large).

When the sample size is small, Fisher's exact test can be used.

These are not part of the course. See textbook Sections 9.1 and 9.2.

Outline

1. Inference for a proportion

2. Chi-squared test
   - Multinomial

Multinomial case

We generalize the previous set up: suppose a population is divided into m categories, with proportions $p_1, p_2, \ldots, p_m$ (so $\sum_i p_i = 1$).

We wish to test for $H_0 : p_1 = p'_1,\, p_2 = p'_2, \ldots,\, p_m = p'_m$.

In a sample of size n, let $n_i \in \mathbb{N}$ be the *observed counts* for the ith category, and let $e_i = n\, p'_i$ be the *expected counts*.

Pearson's **chi-squared** statistic is defined as

$$\chi^2 = \sum_{i=1}^{m} \frac{(n_i - e_i)^2}{e_i},$$

which can be shown, if H_0 is true and $n \to \infty$, to be a χ^2_{m-1} random variable.

Thus, we reject H_0 at significance level α if

$$\chi^2 > \chi^2_{m-1,\,1-\alpha}.$$

Multinomial case, example

In other words, compute

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

and if it is ‘too big’ (bigger than $\chi^2_{m-1,\,1-\alpha}$), then reject the null.

Example – Mendel’s peas

4 characteristic of peas: **S**mooth, **w**rinkled; **Y**ellow, **g**reen.
Offspring of pure **Sw**: $3/4$ **S**, $1/4$ **w**.
Offspring of pure **Yg**: $3/4$ **Y**, $1/4$ **g**.
Their offspring: $9/16$ **SY**, $3/16$ **Sg**, $3/16$ **wY**,$1/16$ **wg**.

Type	**SY**	**Sg**	**wY**	**wg**
Count	315	108	102	31

$m = 4,\ n = 556,\ \chi^2 = 0.604,\ \chi^2_{3,\,0.95} = 7.814.$

Some intuition for the chi-squared test

It is not easy to prove that the χ^2 statistic approximately follows a chi-squared distribution.

Here is a verification for the $m = 2$ case: let X_1 and X_2 denote the random variables for the number of observed in each category. Note that $X_1 + X_2 = n$, and $p_1 + p_2 = 1$.

Then

$$\begin{aligned}\chi^2 &= \frac{(X_1 - p_1 n)^2}{p_1 n} + \frac{(X_2 - p_2 n)^2}{p_2 n} \\ &= \frac{(X_1 - p_1 n)^2}{p_1 n} + \frac{(X_1 - p_1 n)^2}{(1 - p_1)n} \\ &= \frac{(X_1 - p_1 n)^2}{p_1(1 - p_1)n} = \left(\frac{X_1 - p_1 n}{\sqrt{p_1(1 - p_1)n}}\right)^2,\end{aligned}$$

which is the square of one approximately standard normal random variable, i. e. χ^2_1.

Multinomial case, exercise

Exercise – investigate whether the digits of π are random

If they are random, then they would be uniformly distributed. Look at the first 10^{12} digits. Use $\alpha = 0.05$.

Digit	Occurrences
0	99999485134
1	99999945664
2	100000480057
3	99999787805
4	100000357857
5	99999671008
6	99999807503
7	99999818723
8	100000791469
9	99999854780