STAT5002 Weekly Independent Exercises - solution

Sheet 5 - Week 8

STAT5002

1 Tests for proportions

A market research company conducts regular opinion polls to get some idea of support for different political parties over time. On 13 May, 2024 it released the results of a poll based on "a representative cross-section of 1,654 Australian electors from May 6-12, 2024"; see original article. It found that on a "two-party preferred'' basis, 52% of those surveyed indicated support for Labor, while the remaining 48% supported the Liberal/Nationals coalition.

1.1

Describe a box model we could use for making an inference about the "true'' proportion p of electors in Australia who would support Labor on a two-party preferred basis (that is, if asked to choose between Labor and the Liberal/National coalition) based on this data. Comment on the suitability of your model for this example.

Answer:

We would model the data as $n = 1654$ random draws $X_1, \dots, X_n$ made with replacement from a box containing N (equally likely) tickets, with pN marked $\boxed{1}$ and the remaining $(1-p)N$ marked $\boxed{0}$. **Note:** the actual poll used a sample taken **without replacement**, but since the sample size $n = 1,654$ is so small compared to the actual population of ≈ 18 million voters (see the AEC enrolment statistics website), it is OK to ignore the difference between sampling with and without replacement. Aside from this, we can only hope that the company's method for obtaining a "representative sample'' is suitably random.

1.2

Estimate the standard error of the sample proportion. If necessary, round to 3 decimal places.

Answer:

Whatever the true p might be, the sample proportion $\bar{X}$ has $E(\bar{X}) = p$ and $SE(\bar{X}) = \sqrt{\frac{p(1-p)}{n}}$. We can thus provide an estimate of the standard error by "plugging in'' the estimate of p, giving

$$\sqrt{\frac{0.52 * 0.48}{1654}} \approx 0.0123\,,$$

approximately 1.2%.

1.3

Is the value $p = 0.50$ included in a 95% Wilson confidence interval for p based on this data? The R output below may be helpful:

```
round(qnorm(0.95), 2)
```

```
[1] 1.64
```

```
round(qnorm(0.975), 2)
```

```
[1] 1.96
```

Answer:

Note firstly that using the R output, a single random draw from a "standard normal box'' would take a value

- *less* than $+1.96$ with probability 0.975;
- *greater* than $+1.96$ with probability $1 - 0.975 = 0.025$;
- *less* than -1.96 with probability 0.025 (by symmetry) and thus
- between ± 1.96 with probability $1 - 0.025 - 0.25 = 0.95$.

A 95% Wilson confidence interval for p based on this data is given by the set

$$\left\{p\colon |0.52 - p| \le 1.96\sqrt{\frac{p(1-p)}{n}}\right\}.$$

Note that this also the set of values p_0 such that a *two-sided* Z-test of $H_0\colon p = p_0$ would *not be rejected* at the 5% level of significance, i.e. that the corresponding Z-statistic satisfies

$$-1.96 \le \frac{0.52 - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \le 1.96\,.$$

It is also the set of values p such that the observed value $\bar{x} = 0.52$ lies inside the 95% prediction interval for $\bar{X}$, assuming p is the true value, given by $p \pm 1.96\sqrt{\frac{p(1-p)}{n}}$.

Note: we use capital $\bar{X}$ when referring to the random quantity before it is observed, while lower-case $\bar{x}$ is the non-random value we obtain after observing the data.

The Z-statistic for testing $H_0\colon p = 0.5$ takes the value

$$\frac{0.52 - 0.5}{\sqrt{\frac{0.5\times 0.5}{1654}}} = \frac{0.02 \times \sqrt{1654}}{0.5} \approx 1.627\,.$$

Since this is less than 1.96 (in absolute value), we conclude that the data is *consistent* with the (candidate) true value $p = 0.5$ (in this sense), so the value 0.5 is indeed included in the 95% Wilson confidence interval for p based on this data.

1.4

Would the test of $H_0\colon p = 0.5$ versus the alternative $H_1\colon p \neq 0.5$ be rejected at the 5% level of significance? What about at the 10% level of significance?

`Answer:`

We already know that we do not reject H_0 (against the two-sided alternative) at the 5% level of significance, because the value $p = 0.5$ is included in the Wilson 95% confidence interval for p (based on this data).

For the 10% level, note that the two-sided test rejects if the Z-statistic exceeds 1.645 in absolute value (since `qnorm(0.95)` ≈ 1.645).

Since the Z-statistic for testing this H_0 takes the value $1.627 < 1.645$ we do **not** reject at the 10% level either.