STAT5002 Lab7 Solution Sheet

Introduction to Statistics

STAT5002

Between 12 September and 7 November 2017 the Australian Government conducted a national postal survey in order to gauge the Australian community's opinion on changing legislation about same-sex marriage.

In the lead-up to the postal survey, various organisations conducted their own polls to infer how the final survey might turn out. One such poll returned the following results:

In favour of proposed change	Against proposed change	Don't know
1036	581	200

What can we infer from this about the final results of the official postal survey?

1 Simplifying assumptions

Since the final postal survey was to be voluntary, and we ultimately wish to estimate the proportion of participants who favour the proposed change, it is reasonable to

1. discard the "Don't know" counts;
2. model the remaining responses as being like a random sample taken with replacement from a box containing only 0s and 1s, with proportion of 1s equal to p (for some unknown $0 < p < 1$).

The parameter p is then the unknown "population" proportion of voters in favour of the change. If $p > 0.5$, the proposed change to the law will be made by the Government.

Question: Do these assumptions seem reasonable? Comment.

Solution: There are a few points one could mention:

1. The poll might be conducted by sampling **without** replacement, so it might be better if our model used sampling without replacement too. However, since the sample size is so much smaller than the population size, the difference is negligible.
2. The "Don't Know" people may end up voting on the day. If none of them did, our assumptions would be reasonable, but as it is, this is a possible issue with our model.
3. We have no idea how "random" the sampling procedure might have been.

2 Sample size and observed sample proportion

- **What is the sample size n?**

```
# write code here
n = 1036 + 581
n
```

```
[1] 1617
```

- **What is the observed sample proportion $\bar{x}$?** The sample sum is

```
# write code here
s = 1036
s
```

```
[1] 1036
```

so the observed sample proportion is

```
# write code here
xbar = s/n
xbar
```

```
[1] 0.6406926
```

3 Confidence interval

- **Determine a Wilson's 99% confidence interval for p based on the observations in the previous part. Hint:** use `binom.confint()`.

```
# write code here
require(binom)  # you might need to *install* the binom package first, this command *loads* i
```

```
Loading required package: binom
```

```
CI = binom.confint(s, n, conf.level = 0.99, method = "wilson")
CI
```

```
  method    x    n      mean     lower     upper
1 wilson 1036 1617 0.6406926 0.6094411 0.6707943
```

- **Perform a "sanity check" that the endpoints of the interval are such that the observed value $\bar{x}$ is "right on the edge" of a corresponding 99% prediction interval.**

The appropriate multiplier for a 99% prediction interval is 2.576:

```
# write code here
qnorm(0.995)
```

```
[1] 2.575829
```

Note that the endpoints of the Wilson interval are avaiable via the objects `CI$lower` and `CI$upper`:

```
CI$lower
```

```
[1] 0.6094411
```

```
CI$upper
```

```
[1] 0.6707943
```

A 99% prediction interval using the lower endpoint of the Wilson interval is then

```
# write code here
CI$lower + c(-1, 1) * 2.576 * sqrt(CI$lower * (1 - CI$lower)/n)
```

```
[1] 0.5781875 0.6406947
```

Note the upper endpoint of this coincides with our observed $\bar{x}$.

A 99% prediction interval using the upper endpoint of the Wilson interval is then

```
# write code here
CI$upper + c(-1, 1) * 2.576 * sqrt(CI$upper * (1 - CI$upper)/n)
```

```
[1] 0.6406906 0.7008980
```

Note the *lower* endpoint of this coincides with our observed $\bar{x}$.

4 Hypothesis test

The observed proportion is greater than 0.5, but is it significantly greater? Answer this question with an appropriate formal hypothesis test.

Solution:

- **Assumptions**: poll sample is like a random sample taken with replacement from a box containing only 0s and 1s, with proportion of 1s equal to some unknown p.
- **Null Hypothesis**: $H_0 : p = 0.5$
- **Alternative Hypothesis**: $H_1 : 0 > 0.5$ (this is a one-sided test, as the "question" uses the "significantly greater" keywords). This assumes the poll was expecting over 50% in favour, before the data was collected.
- **Test Statistic**: :Larger values of the sample proportion $\bar{X}$ constitute more evidence against H_0, equivalently larger values of the Z-statistic

$$Z = \frac{\bar{X} - E_0(\bar{X})}{SE_0(\bar{X})} = \frac{\bar{X} - 0.5}{\sqrt{\frac{0.5 \times 0.5}{n}}} = 2\sqrt{1617}(\bar{X} - 0.5)\,.$$

- **P-value**: if the observed sample proportion is $\bar{x}$, so the observed value of the Z-statistic is

$$z = 2\sqrt{1617}(\bar{x} - 0.5)$$

  the P-value will be given by `pnorm(z, lower.tail=F)`.

The observed value $\bar{x}$ is

```r
# write any code here
xbar
```

```
[1] 0.6406926
```

so the observed value z is

```r
z = 2 * sqrt(1617) * (xbar - 0.5)
z
```

```
[1] 11.31505
```

and the P-value is

```r
pnorm(z, lower.tail = F)
```

```
[1] 5.527113e-30
```

which is **super-small**!

- **Conclusion** The P-value is very small (much smaller than the significance level of 0.05), providing strong evidence against the null hypothesis of $p = 0.5$. So the observed sample proportion is indeed significantly greater than 0.5.

5 Results

The final outcome of the postal survey (i.e. the true population parameter p) was 61.62% support.

- **In light of the results of your calculations above, what can you say about the Essential poll?**

Solution

The "true" value was indeed included in the 99% Wilson confidence interval for p (although, only just!). So at the same time, the estimate of support is a little high compared to the ultimate "true" value, e.g. a 95% confidence interval:

```r
binom.confint(1036, 1617, conf.level = 0.95, method = "wilson")
```

```
  method    x    n      mean     lower     upper
1 wilson 1036 1617 0.6406926 0.6169988 0.6637196
```

does **not** include (only just though) the ultimate "true" value 0.6162.

. This could be for many reasons, e.g. - poll took a biased sample - the true proportion varied in the weeks/months leading up to the final survey.

Note: you can see the results of the national survey electorate-by-electorate in the following ABC Report.