# STAT2402: Week 4 Computer Laboratory

Inference for Proportions

---

This laboratory session covers the following topics:

1. Hypothesis test for population proportion.
2. Sampling distribution of the sample proportion.
3. Confidence intervals for proportion.

---

1. According to the Center for Disease Control (CDC), the percent of overweight adults 20 years of age and over in the United States is 41.9% (see http://www.cdc.gov/nchs/fastats/obesity-overweight.htm). A city council believes the proportion of overweight citizens in their city is less than this known national proportion. They take a random sample of 150 adults 20 years of age or older in their city and find that 55 are classified as overweight.

   (a) Define the random variable of interest, and state its distribution.
   Solution:
   Let $X$ denote the number of overweight people (out of 150). Then $X \sim \text{Binomial}(150, p)$, where $p$ is the true proportion of overweight people in the city.

   (b) Write down null and alternative hypotheses to be tested corresponding to the question "Is the true proportion of overweight people in the city less than the national proportion?".
   Solution:
   $H_0 : p = 0.419 \qquad H_1 : \; p < 0.419$

   (c) State the distribution of the random variable defined in (a), under $\mathrm{H}_0$.
   Solution:
   If $\mathrm{H}_0$ is true, then $p = 0.0.419$, in which case, $X \sim \text{Binomial}(150, 0.419)$.

   (d) Write down an expression for the $p$–value obtained from the data, and evaluate it using `R`.
   Solution:

$$p\text{–value} = \mathrm{P}(X \leq 55 \mid \mathrm{H}_0) = \mathrm{P}(X \leq 55 \mid p = 0.419) = 0.1115(4 \text{ d.p.; } \texttt{R}).$$

```
pbinom(55, 150, 0.419)

## [1] 0.1115
```

   (e) We can perform the hypothesis test in `R`. The code is
   `binom.test(x, n, p, alternative="greater")`
   where

   - $x$ is the number of success observed
   - $n$ is the number of trials (or sample size)
   - $p$ is the value of the proportion in the null hypothesis
   - alternative states the alternative hypothesis, taking values "greater", "less" or blank for two-sided test. Perform the test and compare your p-value with that obtained kn the previous part. Note the confidence interval in the output.

   Solution:

$$p\text{–value} = \mathrm{P}(X \leq 55 \mid \mathrm{H}_0) = \mathrm{P}(X \leq 55 \mid p = 0.419) = 0.1115(4 \text{ d.p.; } \texttt{R}).$$

```
binom.test(55, 150, 0.419, "less")
```

```
##
##  Exact binomial test
##
## data:  55 and 150
## number of successes = 55, number of trials = 150, p-value
## = 0.1
## alternative hypothesis: true probability of success is less than 0.419
## 95 percent confidence interval:
##  0.0000 0.4363
## sample estimates:
## probability of success
##                 0.3667
```

The same p-value is obtained using the binomial test. The 95% confidence interval for the proportion is $(0, 0.4363)$.

(f) What should the council conclude at the 2.5% level of significance?

Solution:

Since $p$–value $= 0.1115 > 0.025$, there is insufficient evidence to reject the null hypothesis. We conclude based on this analysis that the proportion of overweight people in the city is not less than the national proportion.

(g) Calculate an exact 95% confidence interval for the proportion of obese adults above 20.

Solution:

```
library(GenBinomApps)
clopper.pearson.ci(k = 55, n = 150, alpha = 0.05, CI = "two.sided")

##  Confidence.Interval Lower.limit Upper.limit alpha
##            two.sided      0.2896      0.4492  0.05
```

(h) Also calculate a 95% confidence interval based on the normal distribution. This can be done by using the command `prop.test`. Check the syntax online.

Solution:

```
prop.test(55, 150, p = 0.419)

##
##  1-sample proportions test with continuity correction
##
## data:  55 out of 150, null probability 0.419
## X-squared = 1.5, df = 1, p-value = 0.2
## alternative hypothesis: true p is not equal to 0.419
## 95 percent confidence interval:
##  0.2907 0.4496
## sample estimates:
##      p
## 0.3667
```

(i) Compare the three intervals and explain the differences.

Solution:

The first interval based on the `binom.test` is a one-sided interval, since the hypothesis test is one-sided. If a two-sided hypothesis test is conducted then the interval is:

```
binom.test(55, 150, p = 0.419, alternative = "two.sided")

##
##  Exact binomial test
##
## data:  55 and 150
## number of successes = 55, number of trials = 150, p-value
## = 0.2
## alternative hypothesis: true probability of success is not equal to 0.419
## 95 percent confidence interval:
##  0.2896 0.4492
```

```
## sample estimates:
## probability of success
##                 0.3667
```

This is the same as the exact confidence interval. The interval based on the normal distribution is offset slightly to the left, and is slightly wider.

2. Hypersensitivity of teeth, known as dentin hypersensitivity, is a pathological condition in which teeth are sensitive to thermal, chemical and physical stimuli. Patients with dentin hypersensitivity experience pain from hot/cold and sweet/sour solutions and foods. Pain may also be felt when hot or cold air comes in contact with teeth. Pain varies from mild to excruciating. Dentin hypersensitivity is caused by exposure of dentile tubules from attrition, abrasion, erosion, fracture or chipping of teeth, or a faulty restoration (Kishore, A, Mehrota, K. K. and Saimbi, C. S. (2002) Effectiveness of desensitising agents. *J. of Endodontics*, **28**, 34–35.). Past studies have shown that 5% potassium nitrate solution reduces dentin hypersensitivity in 48% of cases. A researcher is testing the effectiveness of a 40% formalin solution as an alternative to potassium nitrate in reducing hypersensitivity. In a sample of 81 patients suffering from dentin hypersensitivity, 49 reported significant pain relief when using 40% formalin solution. What should the researcher conclude regarding the effectiveness of the two desensitising agents?

   (a) Define the random variable of interest, and state its distribution.

   Solution:

   Let $X$ denote the number of people (out of 81) who experience relief when using a 40% formalin solution. Then $X \sim \text{Binomial}(81, p)$, where $p$ is the true proportion of people who experience relief when using a 40% formalin solution.

   (b) What do you conclude regarding the effectiveness of 1 40% formalin solution compared with the 5% nitrate solution? As part of your answer state appropriate hypotheses and state a p-value. Use a significance level of 0.025.

   Solution:

   We test the null hypothesis $\text{H}_0 : p = 0.48$ against the alternative hypothesis $\text{H}_1 : p > 0.48$. If $\text{H}_0$ is true, then $p = 0.48$, in which case, $X \sim \text{Binom}(81, 0.48)$.That is, the *null distribution* of the test statistic is $\text{Binom}(81, 0.48)$ The p-value of the test is

$$
\begin{aligned}
p\text{–value} &= \text{P}(X \geq 49 \mid \text{H}_0) \\
&= \text{P}(X \geq 49 \mid p = 0.48) \\
&= 0.0161 (4 \text{ d.p.; } \texttt{R}).
\end{aligned}
$$

```r
1 - pbinom(48, 81, 0.48)
```

```
## [1] 0.01609
```

Alternatively,

```r
binom.test(49, 81, 0.48, alternative = "greater")
```

```
##
##  Exact binomial test
##
## data:  49 and 81
## number of successes = 49, number of trials = 81, p-value
## = 0.02
## alternative hypothesis: true probability of success is greater than 0.48
## 95 percent confidence interval:
##  0.5076 1.0000
## sample estimates:
## probability of success
##                 0.6049
```

$$\text{p} - \text{value } = 0.0161 < 0.025,$$

so there is sufficient evidence to reject the null hypothesis. We conclude based on the data that the 40% formalin solution is more effective in desensitising teeth than the 5% nitrate solution.

3. **Driving and cell phones** In a survey, 1640 out of 2246 randomly selected adults in the United States admitted using a cell phone while driving (based on data from Zogby International). The claim is that the proportion of adults who use cell phones while driving is more than 72%. Based on this data, what do you conclude about this claim? Also obtain a 95% confidence interval for the proportion of drivers who use a cell phone while driving.

   Solution:

   Let $X$ denote the number of drivers, out of a random sample of 2246, who use cell phones while driving. Then $X \sim \text{Binomial}(2246, p)$, where $p$ is the true proportion of drivers who use cell phones while driving.

   The hypotheses to be tested are

$$\begin{aligned} \mathrm{H}_0: &\quad p = 0.72 \\ \text{against} \quad \mathrm{H}_1: &\quad p > 0.72. \end{aligned}$$

   Using R,

```
binom.test(1640, 2246, 0.72, alternative = "greater")

##
##  Exact binomial test
##
## data:  1640 and 2246
## number of successes = 1640, number of trials = 2246,
## p-value = 0.1
## alternative hypothesis: true probability of success is greater than 0.72
## 95 percent confidence interval:
##  0.7143 1.0000
## sample estimates:
## probability of success
##                0.7302
```

   The $p$–value, from the observed data, is $\mathrm{P}(X \geq 1640 \mid p = 0.72) = 0.1464$ (4 d.p. from R). Since $0.1464 > 0.025$, there is insufficient evidence to reject $\mathrm{H}_0$ in favour of $\mathrm{H}_1$ at the 2.5% significance level. We conclude that the true proportion of drivers who use cell phones while driving is not more than 72%.

```
clopper.pearson.ci(k = 1640, n = 2246, alpha = 0.05, CI = "two.sided")

##  Confidence.Interval Lower.limit Upper.limit alpha
##            two.sided      0.7113      0.7485  0.05
```

4. According to an article in *The Guardian*, 782 out of 2,370 large companies paid no tax in 2019-20.
   https://www.theguardian.com/australia-news/2021/dec/10/one-third-of-big-businesses-in-australia-still-dont-pay-any-tax-five-years-into-ato-crackdown

   (a) Obtain a point estimate of the proportion of large companies that did not pay tax in 2019-20.

   Solution:

$$\hat{p}_{obs} = \frac{782}{2370}$$

```
(p <- 782/2370)

## [1] 0.33
```

   (b) A sample of 1000 large companies is taken this financial year. Assuming the proportion of such companies that do not pay tax has not changed from that estimated by the Guardian, what is the probability that the sample proportion obtained from this sample is greater than 35%?

   Solution:

   The sample proportion has an approximately normal distribution, with mean and variance given as below:

```r
mean <- p
sd <- sqrt(p * (1 - p)/1000)
1 - pnorm(0.35, mean, sd)

## [1] 0.08884
```

(c) Based on survey in *The Guardian*, obtain a 95% confidence interval for the proportion of large companies that did not pay any tax.
Solution:

```r
clopper.pearson.ci(782, 2370, alpha = 0.05, CI = "two.sided")

##  Confidence.Interval Lower.limit Upper.limit alpha
##            two.sided       0.311      0.3493  0.05
```

---

**Reminder: Logging Off:**
When you have finished, close down RStudio. Remember to log off from your computer before leaving.

---