Day 18 - Parametric inference for one-proportion

Learning outcomes

In this section, we explore how to formally make inference about a single proportion by relying on the Central Limit Theorem.

Motivating example

The following is an excerpt from: *Invasion strategies in round goby (Neogobius melanostomus): is bigger really better?*[1]

> Few studies have systematically investigated mid- or long-term temporal changes of biological characteristics in invasive alien species considering the different phases of an invasion. We studied the invasion performance of one of the most invasive species worldwide, the round goby Neogobius melanostomus, from total absence over first occurrence until establishment from 2010 to 2015 in the upper Danube River. After an upstream movement of the invasion front of about 30 river km within four years, the pattern that round goby pioneering populations significantly differ from longer established ones has been confirmed: Pioneering populations at the invasion front comprised more females than males, and adult specimens with a larger body size compared to those at longer inhabited areas. On the population-level, the proportion of juveniles increased with time since invasion. The results of this study provide support for the previously postulated 'bigger is better' and 'individual trait utility' hypotheses explaining invasion success in round goby. Pioneering invaders with their greater exploratory behavior, highly adaptive phenotypic plasticity and increased competitive ability seem to act as prime emperors of new habitats, strongly following and benefiting from man-made river-bank structures.

The initial estimate of the proportion of juveniles in the Danube in 2010/2011 was 0.09. In the 2015 season, 254 of the 1098 gobies were juvenile. Consider the two following questions:

- Is there evidence that the proportion of juveniles in the Danube in 2015 has increased from the 2011 estimate?
- What is the true proportion of gobies in the Danube in 2015 that are juveniles?

How do these questions differ?

[1] Link to the Dryad Digital Repository for this paper.

Theory-based inference for a single proportion (hypothesis testing)

CLT for proportions

If the _____ conditions are satisfied, the sampling distribution of a sample proportion is approximately normally distributed with mean π and standard deviation $\sqrt{\frac{\pi(1-\pi)}{n}}$.

The sampling distribution of $z = \frac{\hat{p}-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$ is approximately standard normal (mean 0 and standard deviation of 1).

Assumptions for inference

- In order for the Central Limit Theorem to apply for a single proportion, we must expect at least 10 successes and at least 10 failures.

- The observations must be independent of one another.

Hypothesis testing

If the Central Limit Theorem conditions are met, the sampling distribution of $\hat{p}$ under the _____ is approximately normally distributed with mean π_0 and standard deviation $\sqrt{\frac{\pi_0(1-\pi_0)}{n}}$. Under the _____, the sampling distribution of $z = \frac{\hat{p}-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$ is approximately normally distributed with mean 0 and standard deviation of 1.

What is the purpose of hypothesis testing?

When conducting a hypothesis test, what do we always assume to be true?

Theory based estimation for a single proportion (confidence intervals)

Confidence intervals

If the Central Limit Theorem conditions are met, we can estimate the mean and standard deviation of the sampling distribution of $\hat{p}$ by substituting $\hat{p}$ for π in the calculation of the mean and standard deviation of the sampling distribution of $\hat{p}$. The resulting estimated standard deviation of $\hat{p}$ is called the _____.

If the CLT conditions are met, a $(1-\alpha)\%$ confidence interval for π is given by

$$\hat{p} \pm z^*_{1-\alpha/2}\text{SE}(\hat{p})$$

$$\hat{p} \pm z^*_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where $z^*_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution. The quantity $z^*_{1-\alpha/2}\text{SE}(\hat{p})$ is called the _____.

Assumptions for inference

- In order for the Central Limit Theorem to apply for a single proportion, we must observe at least 10 successes and at least 10 failures.

- The observations must be independent of one another.

What is the purpose of creating a confidence interval?

When creating a confidence interval, what do we assume to be true?

Working through the motivating example

1. What is the statistic for these data? In words and notation.

```
x <- 254
n <- 1098
phat <- x/n
```

> $\hat{p} = \frac{254}{1098}$: the sample proportion of gobies that are juvenile.

2. What is the parameter for these data? In words and notation.

> π: the true proportion of all gobies in the Danube that are juvenile in 2015.

3. Is the independence assumption met for these data? Explain why or why not.

> Most likely. In order for the independence assumption to be met, whether a goby is juvenile must not impact whether another goby is juvenile. This is probably reasonable, although there may be some spatial dependence if we happen to sample a number of gobies in the same cute little goby family.

4. Are the CLT conditions met for the hypothesis test?

```
n*.09
```

```
[1] 98.82
```

```
n*(1-.09)
```

```
[1] 999.18
```

> Yes, we expect at least 10 successes and at least 10 failures under the null hypothesis.

5. State the null hypothesis in words and notation.

> $H_0 : \pi = .09$; the true proportion of gobies in the Danube that are juvenile in 2015 is 0.09.

6. State the alternative hypothesis in words and notation.

> $H_a : \pi > 0.09$; the true proportion of gobies in the Danube that are juvenile in 2015 is greater than 0.09.

7. Compute the z-statistic for this test.

```
se.phat.h0 <- sqrt(.09*(1-.09)/n)
z <- (phat - .09)/se.phat.h0
z
```

```
[1] 16.36412
```

8. Use the z-statistic to compute a p-value for this test.

```
1 - pnorm(z,  0, 1)
```

```
[1] 0
```

9. Interpret the p-value in the context of the problem (i.e. as a conditional probability).

> There is a very, very small chance ($< .0001$) of observing a sample proportion of juvenile gobies of 0.231, or even greater than that, if the true proportion of gobies that are juvenile in the Danube in 2015 is actually 0.09.

10. Based on the p-value, which hypothesis seems more plausible?

> Definitely the alternative! Based on the data we have observed, it is extremely unlikely that the true proportion of juvenile gobies in the Danube in 2015 is 0.09; it is likely larger than that.

11. Are the CLT conditions met for a confidence interval?

> Yes, both $n\hat{p}$ and $n(1-\hat{p})$ are greater than 10 (i.e. we observe at least 10 success and 10 failures).

12. Calculate a 95% confidence interval for the parameter.

```
se.phat <- sqrt(phat*(1-phat)/n)
phat + c(-1, 1)*qnorm(.975)*se.phat
```

```
[1] 0.2063876 0.2562718
```

13. Interpret the 95% confidence interval in the context of the problem.

> We are 95% confidence that the true proportion of gobies in the Danube that are juvenile in 2015 is between 0.206 and 0.256.

14. To set up these hypotheses, we treated the sample proportion of gobies that were juvenile in the Danube in 2010/2011 as the hypothesized true proportion for our hypothesis test. Brainstorm why that may be problematic.

> It fails to acknowledge the uncertainty in the estimate from 2010/2011! We should treat both as random - coming soon!