Proportion Inference

Modeling a count of successes

Download the section 14.Rmd handout to `STAT240/lecture/sect14-proportion`.

Download the file `chimpanzee.csv` to `STAT240/data`.

Modeling two numeric variables:

- Use a linear regression model
- β_1 is the unknokwn parameter of interest

Now, we will model a binary outcome:

- Based on the binomial
- p is the unknown parameter of interest

The chimpanzee data is based on an Emory University experiment.

- Chimpanzees choose from colored tokens
- One color is **selfish**, the other is **prosocial**
- Different combinations were studied

p is the probability (proportion) of prosocial. More specifically,

- Is $p_{partner}$ greater than 0.5?
- Is $p_{partner}$ the same as $p_{nopartner}$?

Let X be the number of prosocial choices out of $n = 610$ trials. We have model

$$X \sim Binom(p, 610)$$

which relies on the BINS assumptions.

Consider chimpanzee A.

They made 60 prosocial choices out of 90 with a partner. We have $X = 60$, giving point estimate

$$\hat{p} = \frac{X}{n} = \frac{60}{90}$$

What is the error in this estimate? Let's try a simulation.

Recall that the normal is a good approximation to the binomial when $n \times p$ is large.

$$X \dot{\sim} N(np, \sqrt{np(1-p)})$$

$$\hat{p} \dot{\sim} N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

$$\hat{p} \mathrel{\dot\sim} N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

- $E(\hat{p}) = p$
- Error decreases with n
- Error is a function of p

Inference for p is based on the binomial and normal.

- CI critical values
- Null distribution for a hypothesis test

This depends on n and p. Try simulating the probability for chimpanzee D, with no partner.

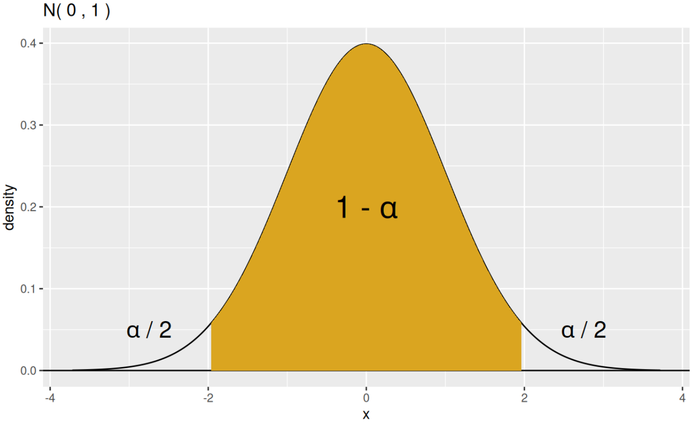
A CI, in general, is

$$\text{point estimate} \;\pm\; \text{critical value} \;\times\; \text{standard error}$$

- For p, the point estimate is $\hat{p}$
- The standard error is $\sqrt{\frac{p(1-p)}{n}}$

The critical value is from $N(0,1)$.

Find a specific quantile with `pnorm`.



For a 90% CI, use 1.645.

Problem: can't find $\sqrt{\frac{p(1-p)}{n}}$. Use $\hat{p}$ instead:

$$\hat{se}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

This is the **Wald** adjustment.

Wald CI for p:

$$\hat{p} \pm z_{\alpha/2} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

We are 90% confident that $p_{A,partner}$ is within (0.585, 0.748).

Another adjustment: add 2 successes and failures to "stabilize" data.

$$n_{AC} = n + 4, \qquad \hat{p}_{AC} = \frac{X + 2}{n + 4}$$

$$\hat{se}(\hat{p}_{AC}) = \sqrt{\frac{\hat{p}_{AC}(1 - \hat{p}_{AC})}{n_{AC}}}$$

This is the **Agresti-Coull** adjustment.

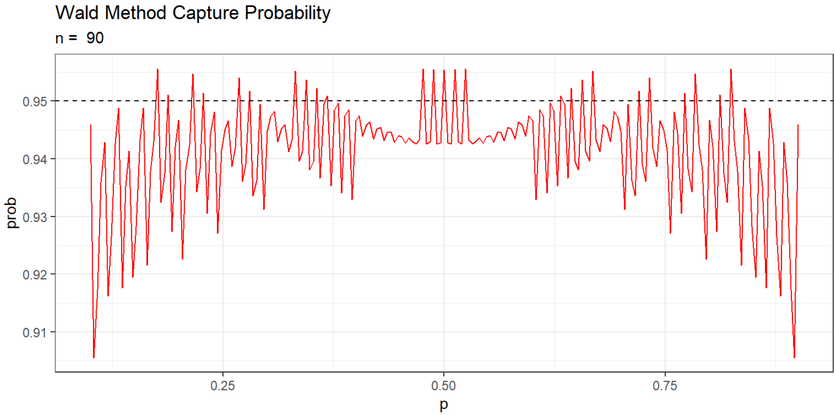
A-C CI for p:

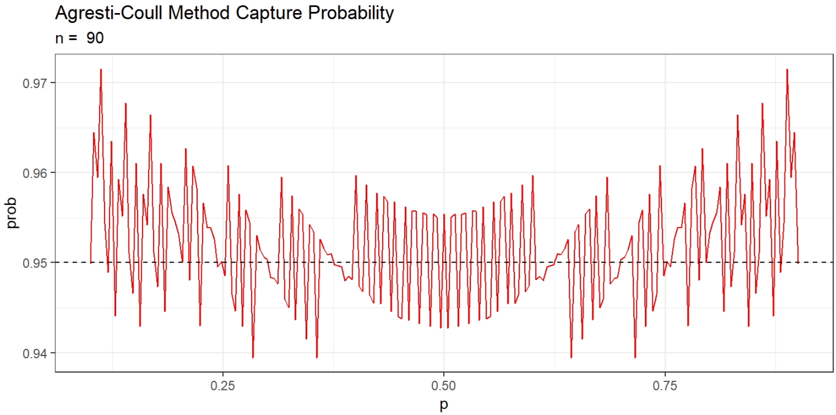
$$\hat{p}_{AC} \pm z_{\alpha/2} \times \sqrt{\frac{\hat{p}_{AC}(1 - \hat{p}_{AC})}{n_{AC}}}$$

We are 90% confident that $p_{A,partner}$ is within (0.579, 0.74).

AC is usually preferable to Wald.

- Both methods are approximate
- AC tends to lead to intervals with coverage probability greater than $1 - \alpha$.


Wald Method Capture Probability
n = 90
prob
0.95
0.94
0.93
0.92
0.91
p
0.25
0.50
0.75


Agresti-Coull Method Capture Probability
n = 90
prob
0.97
0.96
0.95
0.94
0.25
0.50
0.75
p

We model the number of prosocial choices made by chimpanzee A (with a partner) as

$$Binom(90, p)$$

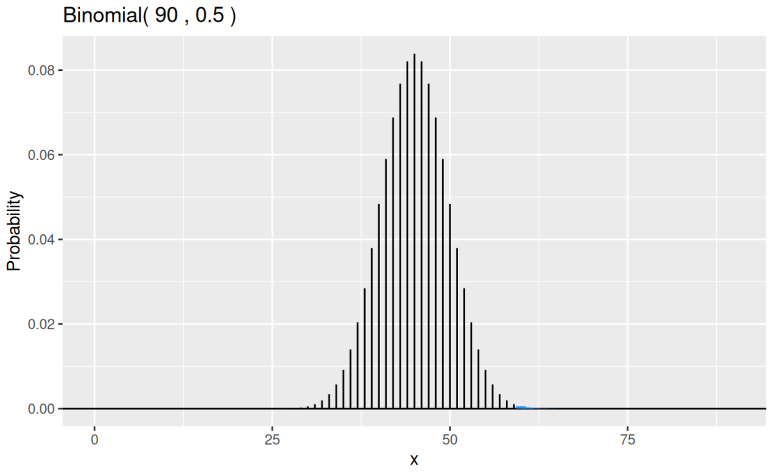
according to the BINS assumptions. Is chimpanzee A more prosocial than selfish?

$$H_0 : p \leq 0.5 \quad \text{versus} \quad H_A : p > 0.5$$

In this context, our test statistic is X (count of prosocial choices) rather than $\hat{p}$. This gives a null distribution of

$$Binom(90, 0.5)$$

and an observed test statistic $x_{obs} = 60$. Values of X higher than 60 give more evidence for H_A.


Binomial(90 , 0.5)
Probability
0.08
0.06
0.04
0.02
0.00
0
25
50
75
x

The p-value is the probability above (and including) $x_{obs} = 60$ on our null distribution.

This is 0.001, which is strong evidence against the null. Chimpanzee A appears to be prosocial more than half the time.

The average prosocial rate with a partner is 0.59. Chimpanzee F has a rate of $47/90$.

Is this significantly *less* than 0.59?

- Set up hypotheses
- Identify null distribution
- Use test statistic $x_{obs} = 47$
- Calculate p-value in the *lower* direction