Week 11 Part II: Confidence Intervals

Yilin

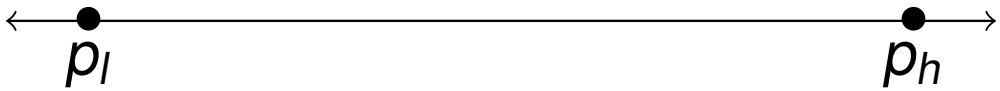
DigiPen

CONFIDENCE INTERVALS

Confidence Intervals

Goal: Determine an interval, $[p_l, p_h]$, that we are "confident" the true value of p lies in. In this case,

- p_l := *lowest* value consistent with the given data
- p_h := *highest* value consistent with the given data
- It is standard to seek a 95% confidence interval, i.e. we are 95% confident that p is in this interval.



We will call $[p_l, p_h]$ the **confidence interval**.

Recall

Previously, we considered the case in which 22 out of 100 individuals tested positive for presence of a toxin. We determined that

$$\hat{p} = \frac{22}{100} = 0.22$$

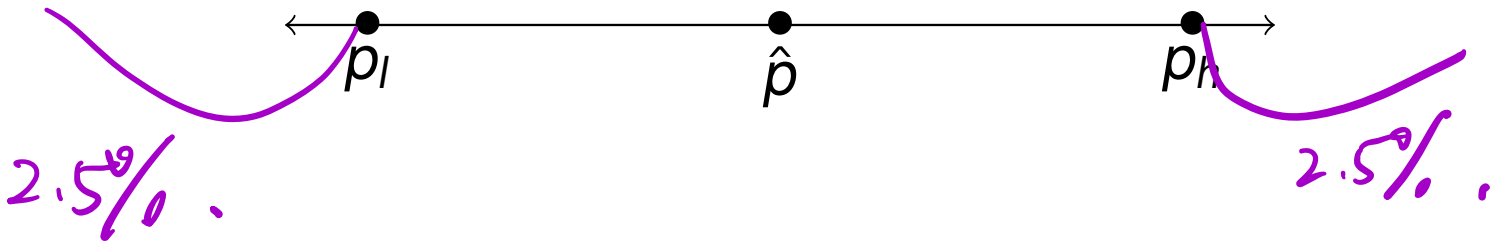
was the **maximum likelihood estimator** for the true value p, where p was the true (but unknown) probability of an individual in the entire population having the toxin present in their body.

New Goal

The goal now is to find an interval $[p_l, p_h]$ containing $\hat{p}$.

- We want to be **95% confident** that the true value of p lies in it.
- We call $[p_l, p_h]$ a **95% confidence interval**.

5% outside $[p_l, p_h]$



Finding p_l and p_h

Question

How do we find p_l and p_h?

- Determine p_l and p_h such that for **any** value of p where $p_l \leq p \leq p_h$, we are "95% confident" that the observed data is "consistent" with our model
- Parameters less than p_l will give the probability the result is **larger** than the data is less than 2.5%.
- Parameters greater than p_h will yield the probability the result is **smaller** than the data is less than 2.5%.

The Method

If X is an r.v. representing a random sample we

- Solve for p_l where

$$\Pr(X \geq \text{ data} \mid p = p_l) = 2.5\%$$

- Solve for p_h where

$$\Pr(X \leq \text{ data} \mid p = p_h) = 2.5\%$$

Example 1: Blood Test

In the blood test example, our **given data** was 22 positive tests out of 100 tested. So if X is the number of positive tests, we find p_l by solving

$$P(X \geq 22 | p = p_l) = 2.5\%,$$

because if the probability that **more** than 22 people have the toxin is 2.5%, then the probability one person has it must be **lower than average ($\hat{p}$)**. Similarly, we find p_h by solving

$$P(X \leq 22 | p = p_h) = 2.5\%,$$

because if the probability that **fewer** than 22 people have the toxin in 2.5%, the probability one person ahs it must be **greater than average ($\hat{p}$)**.

Using the Binomial Distribution

Since we can safely (!) assume that X is a binomial r.v. with $n = 100$ we have

$$P(X \geq 22 \mid p) = \sum_{k=22}^{100} \binom{100}{k} p^k (1-p)^{100-k}$$

So we wish to solve the following polynomial equation for p_l:

$$\sum_{k=22}^{100} \binom{100}{k} p_l^k (1-p_l)^{100-k} = 0.025.$$

Similarly, we find p_h by solving

$$P(X \leq 22 \mid p = p_h) = \sum_{k=0}^{22} \binom{100}{k} p_h^k (1-p_h)^{100-k} = 0.025.$$

Solving

Using a computer we can solve both of these equations and find that

$$p_l \approx 0.1433, \text{ and } p_h \approx 0.3139.$$

Thus the **95% confidence interval** for the parameter p is

$$[0.1433, 0.3139],$$

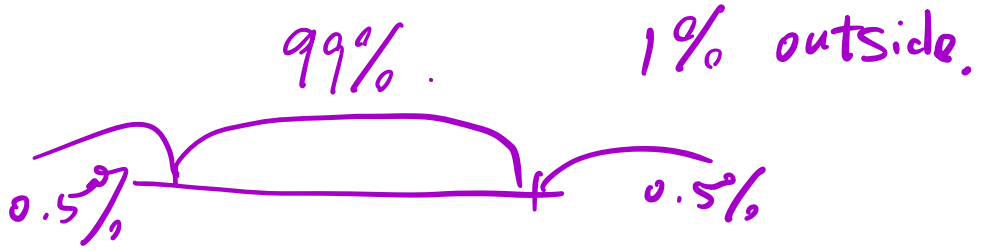
and $\hat{p} = 0.22$ lies in the middle of this interval.

Summary

We are 95% confident that the probability of an individual having the toxin in their blood is between 0.1433 and 0.3139.

Example 2: 99%

With the same experiment, find the 99% confidence interval. Instead of solving for 2.5% we solve for 0.5%.



We find p_l by solving for

$$P(X \geq 22 \mid p_l) = \sum_{k=22}^{100} \binom{100}{k} p_l^k (1 - p_l)^{100-k} = 0.005,$$

and we find p_h by solving for

$$P(X \leq 22 \mid p_h) = \sum_{k=0}^{22} \binom{100}{k} p_h^k (1 - p_h)^{100-k} = 0.005.$$

Solving

Using a computer we can solve these equations and find that

$$p_l \approx 0.124, \text{ and } p_h \approx 0.3437.$$

So the **99% confidence interval** for p is

$$[0.124, 0.3437].$$

Summary

We are 99% confident that the probability of an individual having the toxin in their blood is between 0.124 and 0.3437.

Example 3: Small Sample Size

Suppose five wolves are tested for Lyme disease. Two are found to carry the disease. We seek the MLE (Maximum Likelihood Estimator) $\hat{p}$. Compute

$$L(p) = \Pr(\text{Data} \mid p) = \binom{5}{2} p^2 (1-p)^3 = 10p^2(1-p)^3.$$

To solve for $\hat{p}$, we maximize $L(p)$ using calculus:

$$L'(p) = 20p(1-p)^3 - 30p^2(1-p)^2 = 10p(1-p)^2\,(2(1-p) - 3p)$$

So if $L'(p) = 0$, either $p = 0$, $p = 1$ or

$$0 = 2(1-p) - 3p = 2 - 2p - 3p = 2 - 5p,$$

which has the solution $p = 2/5 = 0.4$. We see that

$$L(0) = 0, \quad L(1) = 0, \quad L(0.4) = 0.3456,$$

so we set $\hat{p} = 0.4$.

Example 2: Confidence

Now we will find the 95% confidence interval. To find p_l, letting X be the random number of diseased wolves out of five, we solve for p_l and p_h where

$$P(X \geq 2 \mid p = p_l) = \sum_{k=2}^{5} \binom{5}{k} p_l^k (1 - p_l)^{5-k} = 0.025,$$

and

$$P(X \leq 2 \mid p = p_h) = \sum_{k=0}^{2} \binom{5}{k} p_h^k (1 - p_h)^{5-k} = 0.025.$$

Solving for p_l and p_h using a computer, we get

$$p_l \approx 0.053, \text{ and } p_h \approx 0.853.$$

So the 95% confidence interval is

$$[0.053, 0.853],$$

which is very large.

Summary

Summary

We are 95% confident that the probability of a wolf being diseased is between 0.053 and 0.853. This is imprecise and the lack of precision is due to the small sample size.

Example 3: Disconnections

Consider a network where 8 users are found to have disconnected. If X is the number of disconnections and we assume X has a Poisson distribution with parameter Λ, then

$$P(X = k) = \frac{\Lambda^k e^{-\Lambda}}{k!}.$$

We found the maximum likelihood estimator for Λ was

$$\hat{\Lambda} = 8.$$

Example 3: Confidence

Now to find the 95% confidence interval we solve for Λ_l and Λ_h where

$$P(X \geq 8 \mid \Lambda_l) = \sum_{k=8}^{\infty} \left(\frac{\Lambda_l^k e^{-\Lambda_l}}{k!} \right) = 0.025,$$

and

$$P(X \leq 8 \mid \Lambda_h) = \sum_{k=0}^{8} \left(\frac{\Lambda_h^k e^{-\Lambda_h}}{k!} \right) = 0.025.$$

Using a computer we find that

$$\Lambda_l = 3.454, \text{ and } \Lambda_h = 15.763.$$

So the confidence interval is

$$[3.454, 15.763]$$

Summary

Summary

We are 95% confident that Λ, the average number of disconnections per user will be between 3.454 and 15.763.