

## Lecture 8: Central Limit Theorem &amp; Binomial Confidence Intervals

Instructor: Ryan Brill

Scribe: Jonathan Pipping

## 8.1 Central Limit Theorem

**Theorem 8.1** (Central Limit Theorem). Suppose  $\{X_i\}_{i=1}^n$  are any collection of **independent and identically-distributed (i.i.d.)** random variables with mean  $\mu = \mathbb{E}[X_i] < \infty$  and variance  $\sigma^2 = \text{Var}[X_i] < \infty$ .

Then as  $n \rightarrow \infty$ , their sum  $S_n = \sum_{i=1}^n X_i$  and sample mean  $\bar{X}_n = \frac{S_n}{n}$  converge to the standard normal distribution. Specifically,

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, 1)$$

$$\frac{\frac{S_n}{n} - \mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow{d} \mathcal{N}(0, 1)$$

**Definition 8.2** (Convergence in Distribution). A sequence of random variables  $\{X_n\}_{n=1}^\infty$  converges in distribution to a random variable  $X$  if for all  $x$  in the support of  $X$ ,

$$\lim_{n \rightarrow \infty} \mathbb{P}(X_n \leq x) = \mathbb{P}(X \leq x)$$

In other words, as  $n \rightarrow \infty$ , the distribution of  $X_n$  converges to the distribution of  $X$ .

What does this mean? If  $Z \sim \mathcal{N}(0, 1)$ , the Central Limit Theorem assures that

$$\mathbb{P}\left(a \leq \frac{S_n - n\mu}{\sigma\sqrt{n}} \leq b\right) \rightarrow \mathbb{P}(a \leq Z \leq b) \text{ as } n \rightarrow \infty$$

Tons of quantities in sports are the sum or mean of i.i.d. random variables, so this normal approximation becomes extremely useful!

## 8.2 The Binomial Parameter

### 8.2.1 Setting Up The Model

Player  $A$  shoots  $n$  free throws, whose results are given by a sequence of random variables  $\{X_i\}_{i=1}^n$ , where

$$X_i = \begin{cases} 1 & \text{if the } i^{\text{th}} \text{ free throw is made} \\ 0 & \text{if the } i^{\text{th}} \text{ free throw is missed} \end{cases}$$

Then  $S_n = \sum_{i=1}^n X_i$  is the total number of free throws made by player  $A$ . We model  $S_n$  as follows:

$$S_n \sim \text{Binomial}(n, p)$$

where  $n$  is the number of free throws attempted and  $p$  is the probability of making a free throw. A Binomial random variable is the sum of  $n$  independent Bernoulli( $p$ ) random variables.

### 8.2.2 Estimating the Binomial Proportion

We want to estimate player  $A$ 's probability of making a free throw,  $p$  from the data  $\{X_i\}_{i=1}^n$ . Our “best guess” of  $p$  is

$$\hat{p} = \frac{S_n}{n} = \frac{1}{n} \sum_{i=1}^n X_i$$

In fact, this is the **Maximum Likelihood Estimator (MLE)** of  $p$ . You will prove this in the Homework. But how confident should we be in this estimate? We answer this question by constructing an asymptotic confidence interval for  $p$ .

### 8.2.3 Confidence Interval for the Binomial Proportion

Recall that each  $X_i$  is a Bernoulli( $p$ ) random variable representing the result of the  $i^{th}$  free throw. We know the mean and variance of a Bernoulli random variable:

$$\begin{aligned}\mu &= \mathbb{E}[X_i] = p \\ \sigma^2 &= \text{Var}[X_i] = p(1-p)\end{aligned}$$

Since  $\{X_i\}_{i=1}^n$  are i.i.d., we can use the CLT to approximate the distribution of  $S_n = \sum_{i=1}^n X_i$ . We don't know the true variance, but we can estimate it by  $\hat{p}(1-\hat{p})$ . Then by the CLT,

$$\frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \xrightarrow{d} \mathcal{N}(0, 1) \text{ as } n \rightarrow \infty$$

Then letting  $z_q$  be the  $q$  quantile (or  $100 \cdot q^{th}$  percentile) of the standard normal distribution, we have

$$\mathbb{P} \left( -z_{1-\alpha/2} \leq \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \leq z_{1-\alpha/2} \right) \approx 1 - \alpha$$

Taking  $\alpha = 0.05$ , we have that  $z_{0.975} \approx 1.96$ . And then

$$\mathbb{P} \left( -1.96 \leq \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \leq 1.96 \right) \approx 0.95$$

Rearranging, we get a 95% confidence interval for  $p$ :

$$\mathbb{P} \left( \hat{p} - 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right) \approx 0.95$$

This is known as the 95% **Wald** Confidence Interval for the Binomial parameter  $p$ . We give a more general formulation below:

**Definition 8.3** (Wald Confidence Interval for  $p$ ). *The  $(1 - \alpha) \cdot 100\%$  Wald Confidence Interval for the Binomial parameter  $p$  is given by*

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

### 8.2.4 M&M's Example

Now suppose you buy an M&M's bag with 56 M&M's, 14 of which are blue. Supposing the color of each M&M is independently drawn from some distribution, what is a 95% confidence interval for  $p_{\text{blue}}$  the probability the company makes an M&M blue?

Here, we have  $n = 56$  and  $\hat{p}_{\text{blue}} = \frac{\text{number of blue M\&M's}}{\text{total number of M\&M's}} = \frac{14}{56} = 0.25$ . Then the 95% Wald Confidence Interval for  $p_{\text{blue}}$  is given by

$$\begin{aligned}\hat{p}_{\text{blue}} \pm z_{0.975} \sqrt{\frac{\hat{p}_{\text{blue}}(1 - \hat{p}_{\text{blue}})}{n}} &= 0.25 \pm 1.96 \sqrt{\frac{0.25(1 - 0.25)}{56}} \\ &= 0.25 \pm 0.1134, \text{ or } [0.1366, 0.3634]\end{aligned}$$

This interval is symmetric about  $\hat{p}_{\text{blue}} = 0.25$  and has length  $2 \times 0.1134 = 0.2268$ : a fairly wide interval! What happens if we were to observe the same proportion of blue M&M's, but with a larger sample size? Let  $n = 400$  and  $\hat{p}_{\text{blue}} = 0.25$ . Then the 95% Wald Confidence Interval for  $p_{\text{blue}}$  is given by

$$\begin{aligned}\hat{p}_{\text{blue}} \pm z_{0.975} \sqrt{\frac{\hat{p}_{\text{blue}}(1 - \hat{p}_{\text{blue}})}{n}} &= 0.25 \pm 1.96 \sqrt{\frac{0.25(1 - 0.25)}{400}} \\ &= 0.25 \pm 0.0424, \text{ or } [0.2076, 0.2924]\end{aligned}$$

This interval is much narrower than the previous one, with a length of only  $2 \times 0.0424 = 0.0848$ . In general, the width of the confidence interval scales with  $O(n^{-1/2})$ . We will now move on to interpreting the confidence interval.

## 8.3 Interpreting the Confidence Interval

### 8.3.1 Frequentist Interpretation

In the previous section, we used the CLT to construct the Wald Confidence Interval in Definition 8.3. We know from this that

$$\mathbb{P}\left(\hat{p} - z_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \approx 1 - \alpha$$

but what does this probability actually mean? Under the model

$$\begin{aligned}S_n &= \sum_{i=1}^n X_i \sim \text{Binomial}(n, p) \\ X_i &\stackrel{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)\end{aligned}$$

$p$  is an **unknown, fixed constant**. This is known as a **frequentist interpretation** of the parameter. What are the implications of this setup? Since  $p$  is fixed, the probability that the confidence interval contains  $p$  is *actually* either 0 or 1. So what does the probability from our confidence interval actually mean?

### 8.3.2 The Confidence Interval as a Random Variable

Believe it or not, **the confidence interval itself is a random variable**. Why? It depends on  $\hat{p}$ , which depends on random variables  $\{X_i\}_{i=1}^n$  through  $S_n$ . So the confidence interval is a random variable, and the

probability that the confidence interval contains  $p$  is the probability that this random interval contains the fixed parameter  $p$ .

If we repeated the experiment many times, in each replication  $p$  remains the same, but the data  $\{X_i\}_{i=1}^n$  changes by randomness, and by extension  $\hat{p}$  and the CI's also change. However, at our specified  $\alpha$  level, we expect the CI to contain  $p$  in  $100(1 - \alpha)\%$  of the replications.

### 8.3.3 Coverage

**Definition 8.4** (Coverage). *The coverage of a confidence interval is the probability that the confidence interval contains the true parameter. Letting  $\theta$  be the true parameter,*

$$\text{Coverage} = \mathbb{P}(\theta \in CI)$$

The Wald Confidence Interval is based on 2 approximations:

1. That  $\mathbb{P}\left(p - z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} \leq p \leq p + z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right) \approx 1 - \alpha$  by the CLT
2. That we can plug in  $\hat{p}$  for  $p$  in  $\sigma = \sqrt{\frac{p(1-p)}{n}}$

Because of these approximations, the **actual** coverage of the  $100(1 - \alpha)\%$  Wald Confidence Interval can be quite far below the **nominal** coverage of  $100(1 - \alpha)\%$  as shown by simulations and computations (Brown, Cai, Dasgupta, 2001). When does this happen?

If  $n$  is several hundred or thousand and/or  $p$  is close to 0.5, the Wald interval is generally **tolerably** accurate. However, if  $n$  is smaller or  $p$  is close to 0 or 1, the Wald interval can be quite inaccurate. To correct for this, Agresti and Coull (1998) recommend introducing **2 artificial successes and failures** into the data before computing  $\hat{p}$ , which is known as the **Agresti-Coull Interval**. We define this interval below.

**Definition 8.5** (Agresti-Coull Interval). *The  $(1 - \alpha) \cdot 100\%$  Agresti-Coull Interval for the Binomial parameter  $p$  is given by*

$$\hat{p} \pm z_{1-\alpha/2} \sqrt{\frac{\hat{p}'(1 - \hat{p}')}{n + 4}}, \text{ where } \hat{p}' = \frac{S_n + 2}{n + 4}$$

In many cases, the Agresti-Coull interval achieves much better coverage than the Wald interval. This is pictured in Figure 8.1, where the Agresti-Coull interval outperforms the Wald interval whenever  $p$  is even moderately far from 0.5.

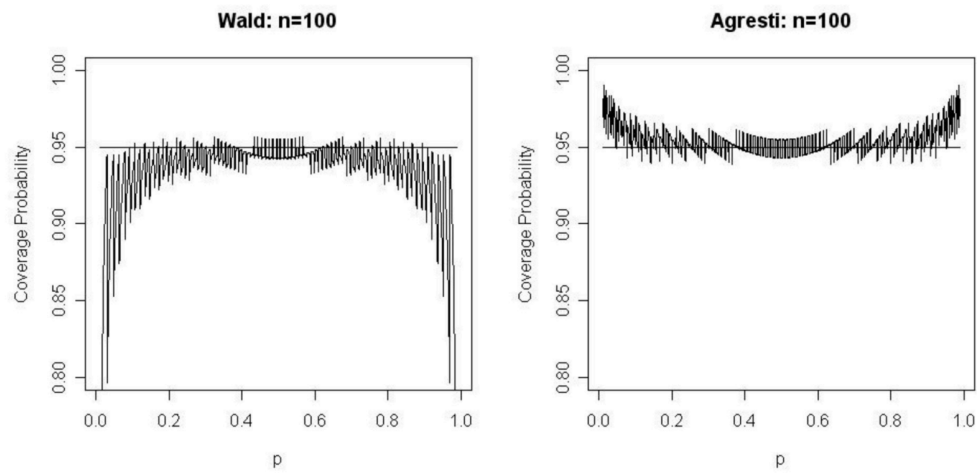


Figure 8.1: Coverage of 95% Wald vs. Agresti-Coull Confidence Intervals

## References

- [AC] Agresti, A., & Coull, B.A., *Approximate is Better than "Exact" for Interval Estimation of Binomial Proportions*, The American Statistician, 1998.
- [BCD] Brown, T.C., Cai, T.T., & Dasgupta, A., *Interval Estimation for a Binomial Proportion*, Statistical Science, 2001.