The Normal Approximation & Binomial Proportion Confidence Interval

Central Limit Theorem

Suppose $\{X_i\}_{i=1}^n$ are any collection of iid random variables with mean $\mu = \mathbb{E}X_i < \infty$ and standard deviation $\sigma = sd(X_i) < \infty$.

Then the sum $S_n = \sum_{i=1}^n X_i$ and mean $\frac{S_n}{n}$ converge in distribution to the normal distribution,

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

$$\frac{\frac{S_n}{n} - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1) \quad \text{as } n \to \infty$$

Convergence in distribution $\xrightarrow{d}$ means

$$\mathbb{P}\left(a \le \frac{S_n - n\mu}{\sigma\sqrt{n}} \le b\right) \to \mathbb{P}(a \le Z \le b) \quad \text{as } n \to \infty.$$

Tons of quantities in sports are the sum or mean of iid random variables! So the normal approximation comes in handy!

Estimating the quality of a free throw shooter

Shaq shoots n free throws, whose results are given by $\{X_i\}_{i=1}^n$, $X_i = \begin{cases} 1 & \text{if } i\text{th free throw made} \\ 0 & \text{if missed} \end{cases}$

$S_n = \sum_{i=1}^n X_i$ is his # made free throws.

We model S_n by

$S_n \sim \text{Binomial}(n,p)$ = # successes in n trials where each trial is independent Bernoulli(p) ($X_i \overset{iid}{\sim} \text{Ber}(p)$)

We want to estimate Shaq's 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 MLE (see the Lab).

But how confident should we be in this guess?

Let $\mu = \mathbb{E}X_i = p$, $\sigma^2 = Var(X_i) = \sqrt{p(1-p)}$.

By CLT, $\frac{\hat{p}-\mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1)$ as $n \to \infty$.

So $\mathbb{P}\left(-2 \le \frac{\hat{p}-\mu}{\sigma/\sqrt{n}} \le 2\right) = \mathbb{P}\left(-2 \le \frac{p-\hat{p}}{\sqrt{p(1-p)/n}} \le 2\right) \approx 0.95$

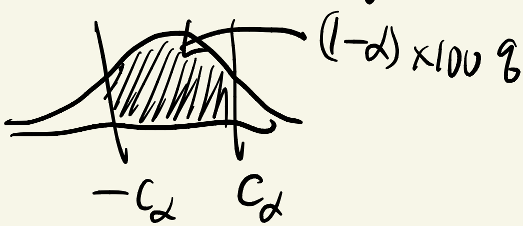
So $p \in \hat{p} \pm 2\sqrt{p(1-p)/n}$ w.p. ≈ 0.95

p is unknown, so use $\hat{p}$ in place of p

Wald CI: $\boxed{\hat{p} \pm 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$

More generally: $\hat{p} \pm C_\alpha \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

is a $(1-\alpha) \times 100\%$ CI where C_α is the quantile of $N(0,1)$ taking $(1-\alpha)\times 100\%$ of the area



You buy an M&M bag with 56 M&M's and 14 blue ones. Supposing the color of each M&M is randomly drawn from some iid distribution, what is a 95% confidence interval of the probability the company makes an M&M blue?

$$\hat{p} = \frac{14}{56} = \frac{1}{4}$$

$$\hat{p} \pm 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \frac{1}{4} \pm 2\sqrt{\frac{\frac{1}{4}\cdot\frac{3}{4}}{n}} = \frac{1}{4} \pm \overset{\approx .85}{\frac{\sqrt{3}}{2}} \cdot \frac{1}{\sqrt{n}}$$

$n = 56 \Rightarrow CI \approx \frac{1}{4} \pm \frac{1}{2}$ pretty wide!

$n = 400 \Rightarrow CI \approx \frac{1}{4} \pm .0425$

The Meaning of a confidence interval

The confidence interval of a Binomial proportion p is

$$\text{Wald CI} = \hat{p} \pm 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

i.e. $\mathbb{P}(p \in CI) \approx 0.95$

but what does this probability $\mathbb{P}$ actually mean?

Under this model, $S_n = \sum_{i=1}^{n} X_i \sim \text{Binomial}(n, p)$

$$X_i \overset{iid}{\sim} \text{Bernoulli}(p)$$

(frequentist)

Under this model, p is an unknown fixed constant.

p either lies in the CI or doesn't lie in it, so in actuality the probability it lies in the CI is either 0 or 1. So, what does $\mathbb{P}$ mean?

The probability $\mathbb{P}$ refers to Randomness in the CI:

The CI itself is a Random variable because $\hat{p}$ is a random variable because $\{X_i\}_{i=1}^{n}$ are Random variables.

If we replicated the experiment 100 times, in each replication p remains the same but the data $\{X_i\}$ and the CI's change by randomness.

on average

the CI is expected to contain the true unobserved p in 95 of these 100 replications

Coverage

The Wald CI is based on 2 approximations:

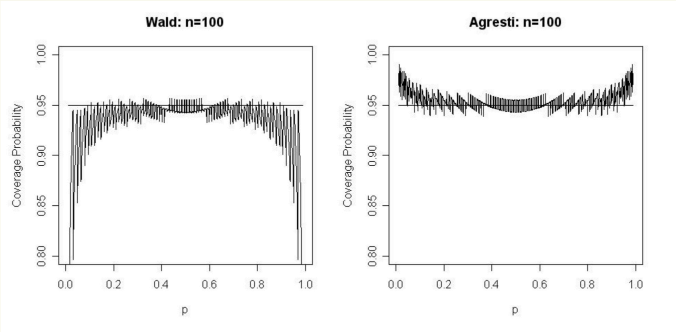
that $\mathbb{P}\left(-2 \le \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \le 2\right) \approx 0.95$ by CLT

and that we can plug in $\hat{p}$ for p in $\sqrt{p(1-p)}$

so that $\mathbb{P}\left(-2 \le \frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}} \le 2\right) \approx 0.95.$

The actual coverage probability $\mathbb{P}(p \in CI)$ of the 95% Wald interval can be quite far below the nominal coverage of 95% as shown by simulations and computations (Brown, Cai, Dasgupta 2001).

If n is several hundred or thousand the Wald interval is tolerably accurate. Otherwise we need to adjust the CI. Agresti and Coull (1998) recommend introducing 2 artificial successes and failures into the data before computing $\hat{p}$ and n, which is better than the Wald interval.



oscillations from the discreteness of Binomial

Agresti Coull CI $= \hat{p}' \pm 2\sqrt{\frac{\hat{p}'(1-\hat{p}')}{n+4}}$ where $\hat{p}' = \frac{S_n+2}{n+4}$.