

Lecture 22 Confidence Interval For The Proportion

BIO210 Biostatistics

Xi Chen

Fall, 2023

School of Life Sciences

Southern University of Science and Technology

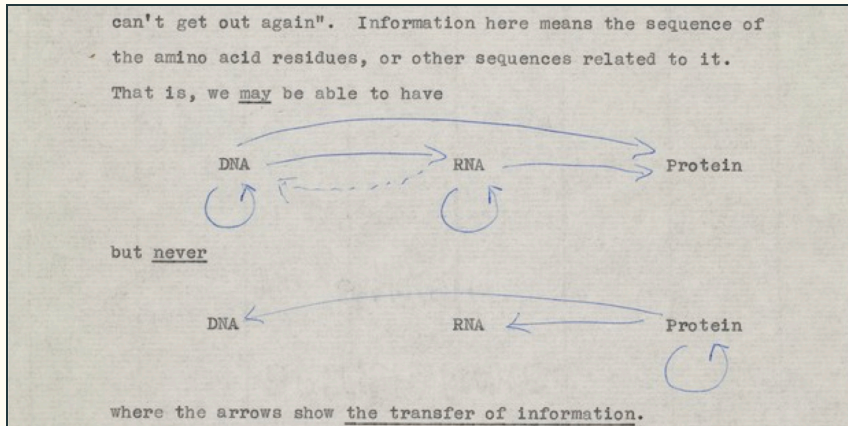


南方科技大学生命科学学院
SUSTech · SCHOOL OF
LIFE SCIENCES

Population Parameters We Have Learnt

Population parameters	Sample statistics
μ	\bar{x}
σ^2	s^2
σ	s
π or p	p or \hat{p}

The Central Dogma



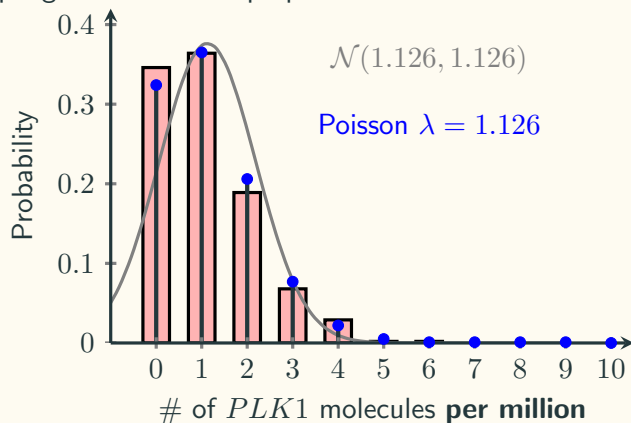
Credit: "Ideas on protein synthesis (Oct. 1956)". Wellcome Collection.

Sample Proportion Example

Gene expression (over-simplified RNA-seq): We know the probability of detecting *PLK1* is $\pi = 0.000001126088083$. If we take a random sample of $n = 1,000,000$ mRNA molecules, what is the sampling distribution of proportion of *PLK1*?

$$\mathcal{N}(\mu = 1.126 \times 10^{-6}, \\ \sigma^2 = 1.126 \times 10^{-12})?$$

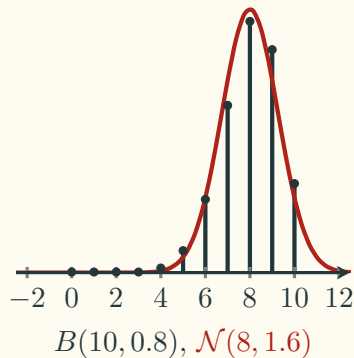
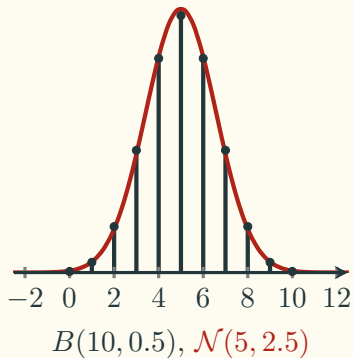
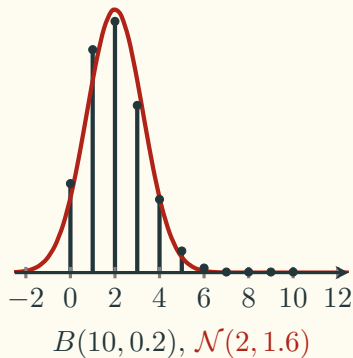
Results from 1,000 samples:
($n = 1,000,000$)



Approximation of The Binomial Distribution

$$B(n, p) \begin{cases} \dot{\sim} \mathcal{N}(\mu = np, \sigma^2 = npq) & , \text{ when } np \geq 10 \text{ and } nq \geq 10 \\ \dot{\sim} \text{Pois}(\lambda = np) & , \text{ when } n \text{ is large, and } p \text{ is small,} \\ & \text{such that } np \text{ is between 0 and 10.} \\ \sim B(n, p) & , \text{ otherwise} \end{cases}$$

The Limitations on np and nq



The Limitations on np and nq

- Binomial: all data are within $[0, n]$
- Normal: no bounds $(-\infty, +\infty)$ for data, but most are within $[\mu - 3\sigma, \mu + 3\sigma]$
- **Intuitively:** when $[\mu - 3\sigma, \mu + 3\sigma]$ is within $[0, n]$, the approximation works well!

$$\mu - 3\sigma > 0$$

$$np - 3\sqrt{npq} > 0$$

$$np > 3\sqrt{npq}$$

$$n^2 p^2 > 9npq$$

$$np > 9q$$

$$np > 9(1 - p) = 9 - 9p$$

$$\mu + 3\sigma < n$$

$$np + 3\sqrt{npq} < n$$

$$n(1 - p) > 3\sqrt{npq}$$

$$n^2 q^2 > 9npq$$

$$nq > 9p$$

$$nq > 9(1 - q) = 9 - 9q$$

Interval Estimation For The Proportion

Goal: for a population containing an unknown proportion (π) of data of our interest, find a and b , such that $\mathbb{P}(a \leq \pi \leq b) = 0.95$.

$$\mathbb{P}(-1.96 \leq Z \leq 1.96) = 0.95$$

$$\mathbb{P}\left(-1.96 \leq \frac{p - \mu_P}{\sigma_P} \leq 1.96\right) = 0.95$$

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

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

Confidence Interval For The Proportion

95% CI For The Sample Proportion

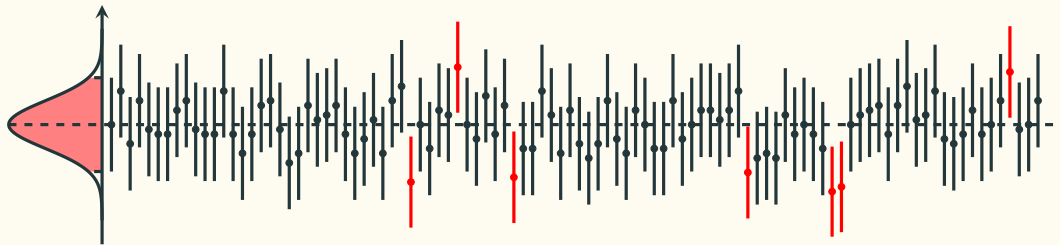
The Wald Interval:

$$\left[p - 1.96\sqrt{\frac{p(1-p)}{n}}, p + 1.96\sqrt{\frac{p(1-p)}{n}} \right]$$

- **Not using t -distribution?** - You don't need to! Remember $\sigma_P = \sqrt{\frac{\pi(1-\pi)}{n}}$, and when p is calculated to estimate π , then σ_P is automatically determined, unlike in the situation of the mean, where you have to do extra (independent) calculation of s to estimate σ , which causes the extra error.

Simulation of 95% CI For The Proportion

100 95% CI for the proportion, constructed using the Wald interval



Probability vs. Statistics

Probability: Previous studies showed that the drug was 80% effective. Then we can anticipate that for a study on 100 patients, on average 80 will be cured and at least 65 will be cured with 99.99% chance.

Statistics: We observe that 78/100 patients were cured by the drug. We will be able to conclude that we are 95% confident that for other studies the drug will be effective on between 69.88% and 86.11% of patients.

Sample Size Estimation Using Confidence Interval of The Proportion

Estimate Sample Size: We want to estimate the percentage of people cured by the drug. Suppose we could draw a truly random sample, and we want a **95% confidence interval estimation** with a **margin of error** no more than $\pm 2\%$. What is the smallest sample size required to obtain the desired margin of error ?

$$95\% \text{ confidence interval: } p \pm 1.96 \sqrt{\frac{p(1-p)}{n}}$$

Goal: find the smallest n such that it **guarantees** that $1.96 \sqrt{\frac{p(1-p)}{n}} \leq 0.02$

Conditions For Interval Estimation For The Proportion

1. Random Samples
2. Normal Condition: the sampling distribution of p needs to be normal
 - $np \geq 10$
 - $nq \geq 10$
3. Independence ($n < 10\%$ population size)

What to do when the normal condition is not met?

- Wilson score interval
- Jeffreys interval
- Agresti–Coull interval
- Arcsine transformation
- Clopper–Pearson interval (the exact method)