Lecture 22 Confidence Interval For The Proportion

BIO210 Biostatistics

Xi Chen

Fall, 2024

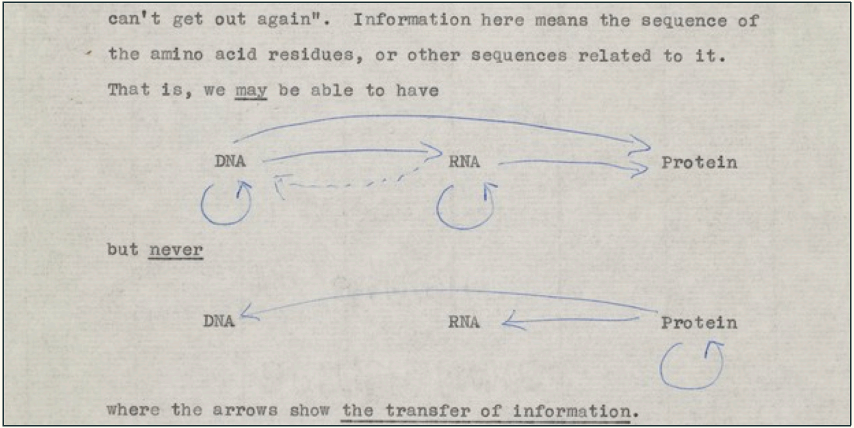
School of Life Sciences
Southern University of Science and Technology

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



can't get out again". Information here means the sequence of the amino acid residues, or other sequences related to it.

That is, we may be able to have

DNA → RNA → Protein

but never

DNA ← RNA ← Protein

where the arrows show the transfer of information.

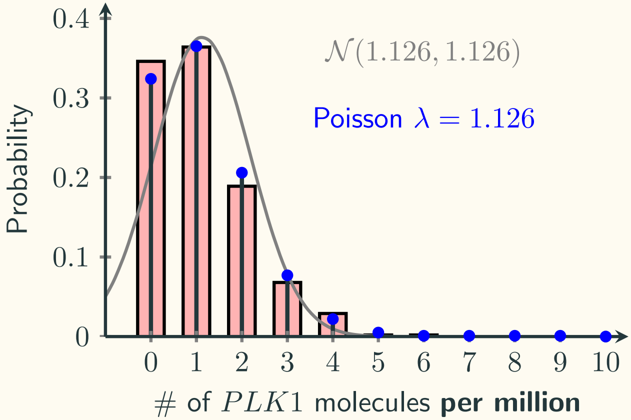
Credit: "Ideas on protein synthesis (Oct. 1956)". Wellcome Collection.

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*?

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

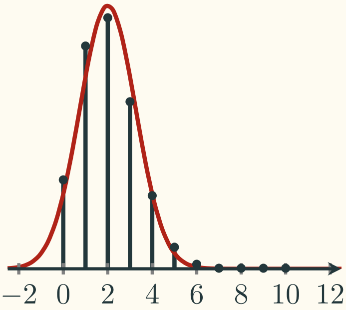
Results from 1,000 samples:
$(n = 1,000,000)$



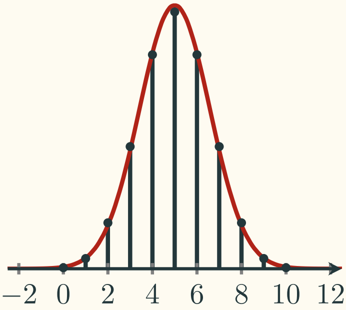
Approximation of The Binomial Distribution

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

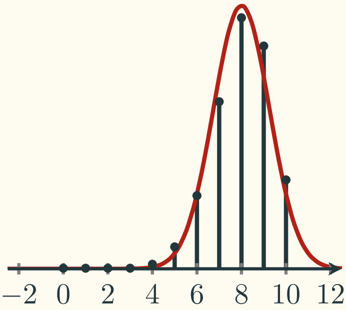
The Limitations on np and nq



$B(10, 0.2)$, $\mathcal{N}(2, 1.6)$



$B(10, 0.5)$, $\mathcal{N}(5, 2.5)$



$B(10, 0.8)$, $\mathcal{N}(8, 1.6)$

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!

$$\begin{aligned} \mu - 3\sigma &> 0 \\ np - 3\sqrt{npq} &> 0 \\ np &> 3\sqrt{npq} \\ n^2p^2 &> 9npq \\ np &> 9q \\ np &> 9(1-p) = 9 - 9p \end{aligned}$$

$$\begin{aligned} \mu + 3\sigma &< n \\ np + 3\sqrt{npq} &< n \\ n(1-p) &> 3\sqrt{npq} \\ n^2q^2 &> 9npq \\ nq &> 9p \\ nq &> 9(1-q) = 9 - 9q \end{aligned}$$

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 $\mathbb{P}(a \leqslant \pi \leqslant b) = 0.95$.

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

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

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

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

Confidence Interval For The Proportion

95% CI For The Sample Proportion

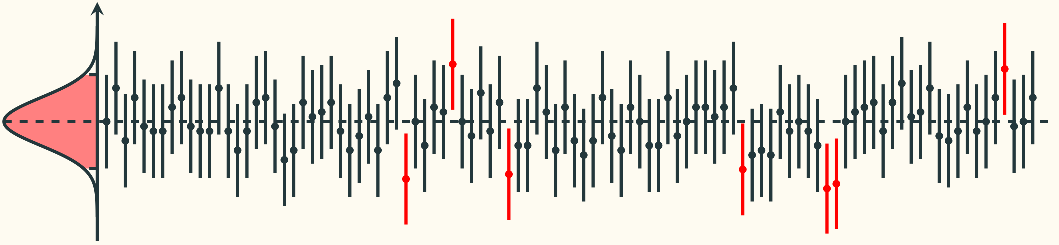
The Wald Interval:

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

- **Not using t-distribution?** - You don't need to! Remember $\sigma_{\boldsymbol{P}} = \sqrt{\frac{\pi(1-\pi)}{n}}$, and when p is calculated to estimate π, then $\sigma_{\boldsymbol{P}}$ 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



An Example in Lecture 1

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 $1.96\sqrt{\frac{p(1-p)}{n}} \leqslant 0.02$

Conditions For Interval Estimation For The Proportion

1. Random Samples

2. Normal Condition: the sampling distribution of p needs to be normal
   - $np \geqslant 10$
   - $nq \geqslant 10$

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)