Recall: CLT:

$X_1 \dots X_n \dots$ iid (independent + identically distrib)

RV and $S_n = \sum_{i=1}^{n} X_i$. Assume $EX_i = \mu$ and

$\mathrm{Var}(X_i) = \sigma^2$. Then $\dfrac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{\text{dist}} N(0,1)$

idea: $S_n \approx N(n\mu, n\sigma^2)$

* Binomial is sum of Bernoulli RV.s so CLT applies.

Q: Half the people support a claim.
A poll asks 1000 people their opinion.
what is P(at least 55% Yes)?

$X = \text{Bin}(1000, \tfrac{1}{2})$. want $P(X \geq 550)$

$$Z = \frac{X - EX}{\sqrt{\text{Var}\, X}} = \frac{X - 500}{\sqrt{1000 \cdot \frac{1}{2} \cdot \frac{1}{2}}} \quad \text{is roughly } N(0,1)$$

$$X \geq 550 \iff \frac{X - 500}{\sqrt{250}} \geq \frac{50}{\sqrt{250}} = 3.18\ldots$$

$$P(X \geq 550) \approx P(N(0,1) \geq 3.18) = 1 - \Phi(3.18) < 0.001$$

---

e.g. Each job takes a server $\text{Exp}(1)$ to finish.

what is $P(50 \text{ jobs take} \geq 60 \text{ time})$?

Sol: X_i = time for job i. Each is $\text{Exp}(1)$

need $P(S_{50} \geq 60)$

note: S_n is $\text{Gamma}(n, 1)$.

Here $\mu = 1$ and $\sigma^2 = 1$

Exact : $$P(S_{50} \geq 60) = \int_{60}^{\infty} \frac{x^{49}}{49!} e^{-x} dx = 0.084$$

CLT : $$P(S_{50} \geq 60) = P\left(\frac{S_{50} - n\mu}{\sqrt{n}\,\sigma} \geq \frac{60 - n\mu}{\sqrt{n}\,\sigma} \right)$$

$$\approx P\left(N(0,1) \geq \underbrace{\frac{60 - n\mu}{\sqrt{n}\,\sigma}} \right) = 1 - \Phi(1.4) = 0.079$$

$$\frac{60 - 50 \cdot 1}{\sqrt{50} \cdot 1} \approx 1.4$$

Statistics

def: Sample mean of samples $X_1 \dots X_n$ is

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

If (X_i) are iid samples from some distrib with expectation μ, then $\bar{X}$ is an estimator for μ.

Claim: $\bar{X}$ is an unbiased estimator: $\mathbb{E}\bar{X} = \mu$

e.g. $(X_i) = \{3, 1, 4, 1, 5, 9\}$ $\quad \bar{X} = \frac{23}{6}$

def: Sample variance is $\delta^2 = \frac{1}{n-1}\sum (X_i - \bar{X})^2$

Claim: δ^2 is unbiased estimator for $\mathrm{Var}(X_i)$

i.e. $E\delta^2 = Var(X_i)$

Proof: Let $S = \sum_{i=1}^{n} x_i$ so $\bar{X} = \frac{S}{n}$.

$$ES = n\mu \quad \text{where} \quad \mu = EX_i$$

$$ES^2 = E\left(\sum x_i^2 + \sum_{i \neq j} x_i x_j\right) = n E(X_i^2) + 2\binom{n}{2} \underbrace{EX_i)(EX_j}_{\mu^2}$$

$$E\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right) = E\left(\sum x_i^2 - \underbrace{2\sum x_i \bar{x}}_{-2S\bar{X}} + \underbrace{n\bar{X}^2}_{n\bar{X}^2}\right)$$

$$= nE(X_i^2) - E\frac{S^2}{n}$$

$$= nE(X_i^2) - E(X_i)^2 - (n-1)\mu^2 = (n-1)\left(E(X^2) - \mu^2\right)$$

$$= (n-1)\,Var(X) \qquad \square$$

Hypothesis testing:

Get some data. Hypothesis: This comes from a given distrib. Determine the likelyhood of data (or better, given hypothesis. (This is p-value).

If p-value small, reject hypothesis

Likelyhood = $P(\text{Data} \mid \text{Hyp})$

e.g. Hyp. Coin is fair.

Data: 550 Heads, 450 Tails in 1000 tosses

$P(\text{Data} \mid \text{Hyp}) < 10^{-3}$

Note: we want p-value = $P(\geq 550 \text{ Heads} \mid \text{Hyp})$

Rule out the Hyp. if $X = \#\text{heads}$ satisfies

$|X - 500| \geq \ell$ where ℓ is such that

$P(|X - 500| \geq \ell \mid \text{fair coin}) = 5\%$ (level of certainty)

For 1000 coins, this ℓ is 1.96 std dev.

So $\ell = 31$