24 2025-03-10 | Week 9 | Lecture 23

Example 76 (Birthday paradox)**.** Let

$$E = [\text{At least one birthday match in a room of } n \text{ people }]$$

Want to compute

$$\mathbb{P}[E].$$

We will use

$$\mathbb{P}[E] = 1 - \mathbb{P}[E^c].$$

In words, $E^c = [\text{no birthday match}]$. Let's assume there are 365 days in a year, and that all days are equally likely. Let

$$D_j = \mathbb{P}[j\text{-th person differs from all predecessors}]$$

Then

$$\begin{aligned}\mathbb{P}[E] &= 1 - \mathbb{P}[E^c] \\ &= 1 - \mathbb{P}[D_1]\,\mathbb{P}[D_2 \mid D_1]\,\mathbb{P}[D_3 \mid D_1 D_2] \ldots \mathbb{P}[D_n \mid D_1 \cdots D_{n-1}] \\ &= 1 - \frac{365}{365} \cdot \frac{364}{365} \cdot \frac{363}{365} \cdots \frac{365-(n-1)}{365}\end{aligned}$$

When $n = 23$, $\mathbb{P}[E] > \frac{1}{2}$. When $n = 35$, $\mathbb{P}[E] \approx .8$.

End of Example 76. □

introduced the central limit theorem

Definition 77 (iid)**.** We say that a sequence $X_1, \ldots, X_n$ of random variables is *independent and identically distributed (iid)* if

- The X_i's are independent rvs
- Every X_i has the same probability distribution as X_1

In statistics, such a sequence of r.v.s is called a *random sample from the distribution X_1* The partial sum is

$$T_n = X_1 + \ldots + X_n$$

The *sample mean* is

$$\overline{X} = \frac{X_1 + \ldots + X_n}{n}$$

and the *sample variance* is $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$.

Recall that $\mathcal{N}(\alpha, \beta)$ means "normal distribution with mean α and variance β".

Theorem 78 (Central Limit Theorem)**.** *Let $X_1, \ldots, X_n$ be a random sample from a distribution with mean μ and variance σ^2. If n is sufficiently large, then*

$$T_n \text{ is approximately } \mathcal{N}\left(n\mu, n\sigma^2\right)$$

and

$$\overline{X} \text{ is approximately } \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

The larger n, the better the approximation. Usually $n = 30$ is big enough.

Example 79. Flip a coin 1000 times. What is the probability that you get at least 550 heads?

Let T be the total number of heads. Then T is binomial distribution with $n = 1000$ and $p = \frac{1}{2}$. For each $k \in \{0, 1, \ldots, 1000\}$, we have the formula

$$\mathbb{P}[T = k] = \binom{n}{k} p^k (1-p)^k = \binom{n}{k} \frac{1}{4^k}$$

So we want

$$\mathbb{P}[T \geq 550] = \sum_{k=550}^{1000} \binom{n}{k} \frac{1}{4^k}$$

Good luck with that. Instead, let's use the central limit theorem. Write

$$T = X_1 + X_2 + \ldots + X_{1000}$$

where

$$X_i = \begin{cases} 0 & : \quad i^{th} \text{ coin flip is tails} \\ 1 & : \quad i^{th} \text{ coin flip is heads} \end{cases}$$

Then

$$\mathbb{E}[X_i] = \frac{1}{2}$$

and

$$\begin{aligned} \mathrm{Var}(X_i) &= \mathbb{E}\left[X_i^2\right] - (\mathbb{E}[X])^2 \\ &= \frac{1}{2} - \left(\frac{1}{2}\right)^2 \\ &= \frac{1}{2} \end{aligned}$$

So by the central limit theorem, T is approximately $\mathcal{N}(1000\mu, 1000\sigma^2)$ where $\mu = \frac{1}{2}$ and $\sigma^2 = \frac{1}{2}$ or

$$T \sim \mathcal{N}(500, 500)$$

Therefore using the standardization trick,

$$\begin{aligned} \mathbb{P}[T \geq 550] &\approx \mathbb{P}\left[Z \geq \frac{550 - 500}{\sqrt{500}}\right] \\ &\approx \mathbb{P}[Z \geq 2.236] \\ &= \frac{1}{\sqrt{2\pi}} \int_{2.236}^{\infty} e^{-x^2/2} dx \\ &\approx 0.013 \end{aligned}$$

End of Example 79. $\square$