Artificial Intelligence and Machine Learning

Barbara Caputo

Parameter estimation: MLE, MAP

Estimating Probabilities

Flipping a Coin

I have a coin, if I flip it, what's the probability it will fall with the head up?

Let us flip it a few times to estimate the probability:

The estimated probability is: **3/5** "Frequency of heads"

Why???... and How good is this estimation???

MLE for Bernoulli distribution

Data, D =

$$D = \{X_i\}_{i=1}^n, \ X_i \in \{\mathrm{H}, \mathrm{T}\}$$

MLE for Bernoulli distribution

Data, D =

$$D = \{X_i\}_{i=1}^n, \; X_i \in \{\mathrm{H}, \mathrm{T}\}$$

$P(Heads) = \theta, \; P(Tails) = 1-\theta$

Flips are **i.i.d.**:

- **Independent** events
  - **Identically distributed** according to Bernoulli distribution

MLE: Choose θ that maximizes the probability of observed data

Maximum Likelihood Estimation

MLE: Choose θ that maximizes the probability of observed data

Maximum Likelihood Estimation

MLE: Choose θ that maximizes the probability of observed data

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} \; P(D \mid \theta)$$

Maximum Likelihood Estimation

MLE: Choose θ that maximizes the probability of observed data

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} P(D \mid \theta)$$

$$= \arg\max_{\theta} \prod_{i=1}^{n} P(X_i|\theta)$$

Independent draws

Maximum Likelihood Estimation

MLE: Choose θ that maximizes the probability of observed data

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} P(D \mid \theta)$$

$$= \arg\max_{\theta} \prod_{i=1}^{n} P(X_i|\theta)$$

Independent draws

$$= \arg\max_{\theta} \prod_{i:X_i=H} \theta \prod_{i:X_i=T} (1-\theta)$$

Identically distributed

Maximum Likelihood Estimation

MLE: Choose θ that maximizes the probability of observed data

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} P(D \mid \theta)$$

$$= \arg\max_{\theta} \prod_{i=1}^{n} P(X_i|\theta)$$ Independent draws

$$= \arg\max_{\theta} \prod_{i:X_i=H} \theta \prod_{i:X_i=T} (1-\theta)$$ Identically distributed

$$= \arg\max_{\theta} \underbrace{\theta^{\alpha_H}(1-\theta)^{\alpha_T}}_{J(\theta)}$$

Maximum Likelihood Estimation

MLE: Choose θ that maximizes the probability of observed data

Maximum Likelihood Estimation

MLE: Choose θ that maximizes the probability of observed data

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} \; P(D \mid \theta)$$

Maximum Likelihood Estimation

MLE: Choose θ that maximizes the probability of observed data

$$
\begin{aligned}
\widehat{\theta}_{MLE} &= \arg\max_{\theta} \quad P(D \mid \theta) \\
&= \arg\max_{\theta} \underbrace{\theta^{\alpha_H}(1-\theta)^{\alpha_T}}_{J(\theta)}
\end{aligned}
$$

Maximum Likelihood Estimation

MLE: Choose θ that maximizes the probability of observed data

$$\begin{aligned}\widehat{\theta}_{MLE} &= \arg\max_{\theta} \; P(D \mid \theta) \\ &= \arg\max_{\theta} \; \underbrace{\theta^{\alpha_H}(1-\theta)^{\alpha_T}}_{J(\theta)}\end{aligned}$$

$$\frac{\partial J(\theta)}{\partial \theta} = \alpha_H \theta^{\alpha_H - 1}(1-\theta)^{\alpha_T} - \alpha_T \theta^{\alpha_H}(1-\theta)^{\alpha_T - 1}\Big|_{\theta=\hat{\theta}_{\mathrm{MLE}}} = 0$$

Maximum Likelihood Estimation

MLE: Choose θ that maximizes the probability of observed data

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} \; P(D \mid \theta)$$

$$= \arg\max_{\theta} \; \underbrace{\theta^{\alpha_H}(1-\theta)^{\alpha_T}}_{J(\theta)}$$

$$\frac{\partial J(\theta)}{\partial \theta} = \alpha_H \theta^{\alpha_H - 1}(1-\theta)^{\alpha_T} - \alpha_T \theta^{\alpha_H}(1-\theta)^{\alpha_T - 1}\Big|_{\theta=\hat{\theta}_{\mathrm{MLE}}} = 0$$

$$\alpha_H(1-\theta) - \alpha_T\theta\Big|_{\theta=\hat{\theta}_{\mathrm{MLE}}} = 0$$

$$\boxed{\widehat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}}$$

What about prior knowledge?

We know the coin is “close” to 50-50. What can we do now?

What about prior knowledge?

We know the coin is “close” to 50-50. What can we do now?

The Bayesian way…

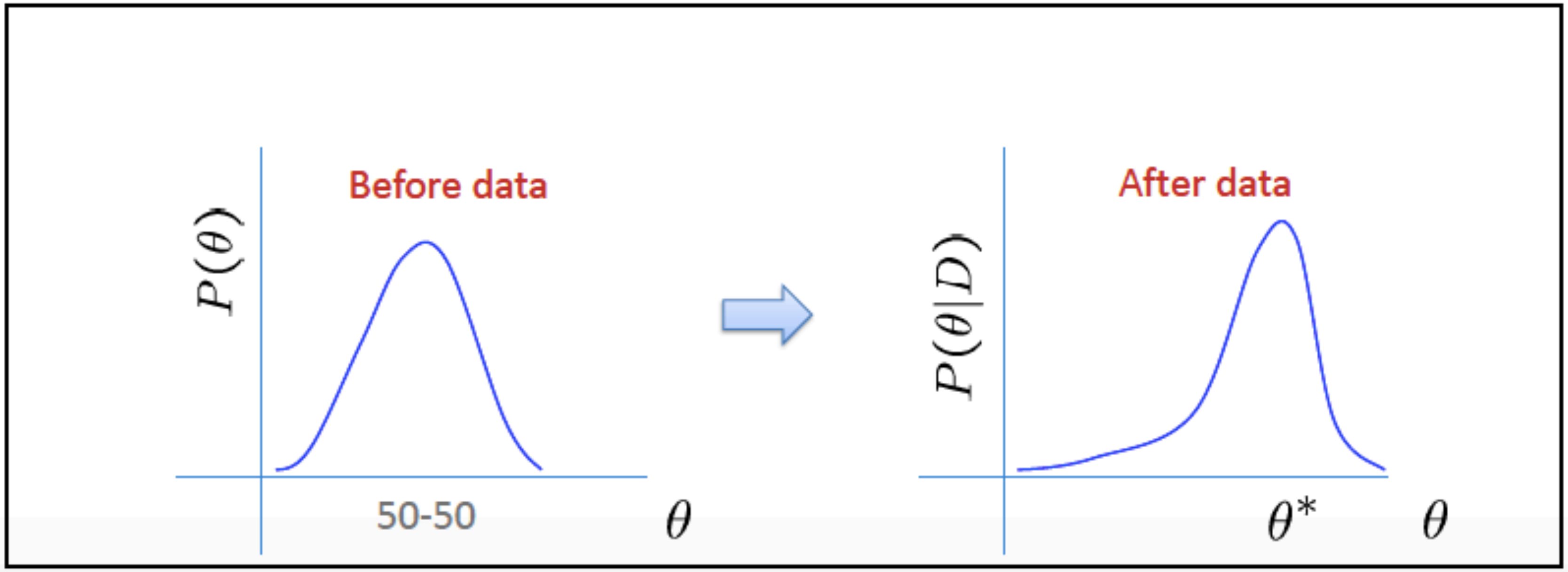
Rather than estimating a single θ, we obtain a distribution over possible values of θ

What about prior knowledge?

We know the coin is “close” to 50-50. What can we do now?

The Bayesian way…

Rather than estimating a single θ, we obtain a distribution over possible values of θ



Bayesian Learning

- Use Bayes rule:

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, **53:370-418**

Bayesian Learning

- Use Bayes rule:

$$P(\theta \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \theta) P(\theta)}{P(\mathcal{D})}$$

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, **53:370-418**

Bayesian Learning

- Use Bayes rule:

$$P(\theta \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \theta)P(\theta)}{P(\mathcal{D})}$$

- Or equivalently:

$$\underbrace{P(\theta \mid \mathcal{D})}_{\text{posterior}} \propto \underbrace{P(\mathcal{D} \mid \theta)}_{\text{likelihood}}\underbrace{P(\theta)}_{\text{prior}}$$

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, **53:370-418**

MAP estimation for Binomial distribution

Coin flip problem

MAP estimation for Binomial distribution

Coin flip problem

Likelihood is Binomial $P(\mathcal{D} \mid \theta) = \binom{n}{\alpha_H} \theta^{\alpha_H}(1-\theta)^{\alpha_T}$

MAP estimation for Binomial distribution

Coin flip problem

Likelihood is Binomial $P(\mathcal{D} \mid \theta) = \binom{n}{\alpha_H} \theta^{\alpha_H}(1-\theta)^{\alpha_T}$

If the prior is Beta distribution,

$$P(\theta) = \frac{\theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} \sim Beta(\beta_H, \beta_T)$$

MAP estimation for Binomial distribution

Coin flip problem

Likelihood is Binomial $P(\mathcal{D} \mid \theta) = \binom{n}{\alpha_H} \theta^{\alpha_H} (1-\theta)^{\alpha_T}$

If the prior is Beta distribution,

$$P(\theta) = \frac{\theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} \sim Beta(\beta_H, \beta_T)$$

$\Rightarrow$ posterior is Beta distribution

$$P(\theta|D) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

P(θ) and P(θ| D) have the same form! [Conjugate prior]

MAP estimation for Binomial distribution

Coin flip problem

Likelihood is Binomial $P(\mathcal{D} \mid \theta) = \binom{n}{\alpha_H} \theta^{\alpha_H} (1-\theta)^{\alpha_T}$

If the prior is Beta distribution,

$$P(\theta) = \frac{\theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} \sim Beta(\beta_H, \beta_T)$$

$\Rightarrow$ posterior is Beta distribution

$$P(\theta|D) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

P(θ) and P(θ| D) have the same form! [Conjugate prior]

$$\widehat{\theta}_{MAP} = \arg\max_{\theta} P(\theta \mid D) = \arg\max_{\theta} P(D \mid \theta) P(\theta) = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + \beta_T - 2}$$

MLE vs. MAP

- Maximum Likelihood estimation (MLE)

  Choose value that maximizes the probability of observed data

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} P(D|\theta)$$

MLE vs. MAP

- Maximum Likelihood estimation (MLE)

  Choose value that maximizes the probability of observed data

$$\hat{\theta}_{MLE} = \arg\max_{\theta} P(D|\theta)$$

- Maximum *a posteriori* (MAP) estimation

  Choose value that is most probable given observed data and prior belief

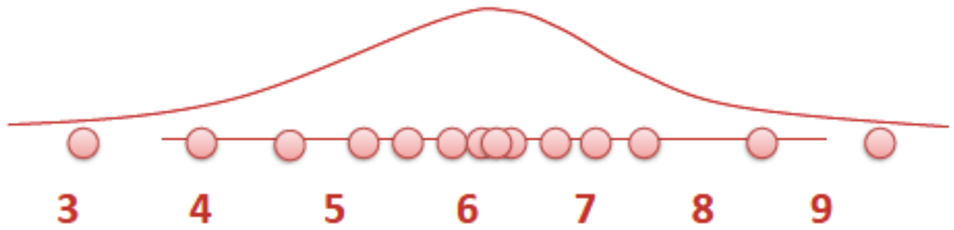
$$\begin{aligned}\hat{\theta}_{MAP} &= \arg\max_{\theta} P(\theta|D) \\ &= \arg\max_{\theta} P(D|\theta)P(\theta)\end{aligned}$$

Bayesians vs.Frequentists

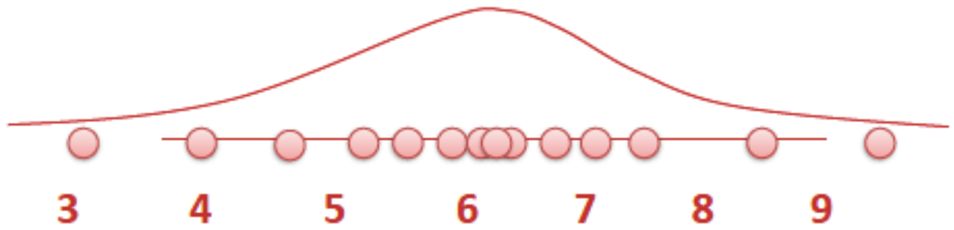
You are no good when sample is small

You give a different answer for different priors

What about continuous features?

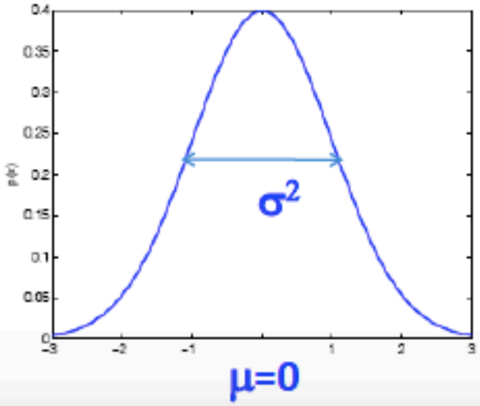


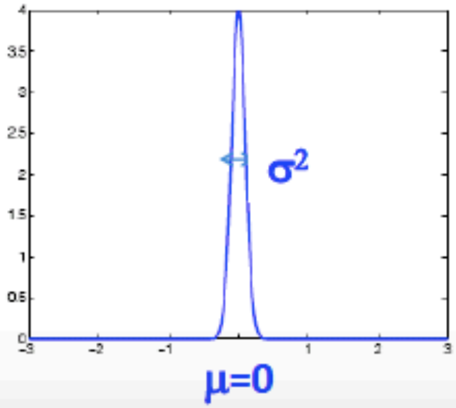
What about continuous features?



Let us try Gaussians…

$$p(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) = \mathcal{N}_x(\mu, \sigma)$$





MLE for Gaussian mean and variance

Choose $\theta = (\mu, \sigma^2)$ that maximizes the probability of observed data

MLE for Gaussian mean and variance

Choose θ= (μ,σ²) that maximizes the probability of observed data

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} \; P(D \mid \theta)$$

MLE for Gaussian mean and variance

Choose θ= (μ,σ^2) that maximizes the probability of observed data

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} P(D \mid \theta)$$

$$= \arg\max_{\theta} \prod_{i=1}^{n} P(X_i|\theta)$$

Independent draws

MLE for Gaussian mean and variance

Choose $\theta = (\mu,\sigma^2)$ that maximizes the probability of observed data

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} \; P(D \mid \theta)$$

$$= \arg\max_{\theta} \prod_{i=1}^{n} P(X_i|\theta)$$

Independent draws

$$= \arg\max_{\theta} \prod_{i=1}^{n} \frac{1}{2\sigma^2} e^{-(X_i-\mu)^2/2\sigma^2}$$

Identically distributed

MLE for Gaussian mean and variance

Choose $\theta = (\mu, \sigma^2)$ that maximizes the probability of observed data

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} P(D \mid \theta)$$

$$= \arg\max_{\theta} \prod_{i=1}^{n} P(X_i|\theta)$$

Independent draws

$$= \arg\max_{\theta} \prod_{i=1}^{n} \frac{1}{2\sigma^2} e^{-(X_i-\mu)^2/2\sigma^2}$$

Identically distributed

$$= \arg \max_{\theta=(\mu,\sigma^2)} \underbrace{\frac{1}{2\sigma^2} e^{-\sum_{i=1}^{n}(X_i-\mu)^2/2\sigma^2}}_{J(\theta)}$$

MLE for Gaussian mean and variance

MLE for Gaussian mean and variance

$$\widehat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\widehat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \widehat{\mu})^2$$

MLE for Gaussian mean and variance

$$\widehat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\widehat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \widehat{\mu})^2$$

Note: MLE for the variance of a Gaussian is **biased**

[Expected result of estimation is **not** the true parameter!]

Unbiased variance estimator: $\widehat{\sigma}^2_{unbiased} = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \widehat{\mu})^2$

Five minutes break!

The Law of Large Numbers*

This is a key fundamental result, which is so intuitive that it is at the heart of the frequentist interpretation of probability, and is essential to the field of statistics.

The Law of Large Numbers*

This is a key fundamental result, which is so intuitive that it is at the heart of the frequentist interpretation of probability, and is essential to the field of statistics.

Theorem: The Law of Large Numbers (LLN)

Let $X_1, X_2, \ldots$ be an infinite sequence of independent and identically distributed (i.i.d.) random variables, with mean $\mu_X = \mathbb{E}[X_i]$. Define the average

$$\overline{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i \ .$$

Then, as $n \to \infty$, the average converges to a non-random real number. Specifically

$$\lim_{n\to\infty} \overline{X}_n = \mu_x \ .$$

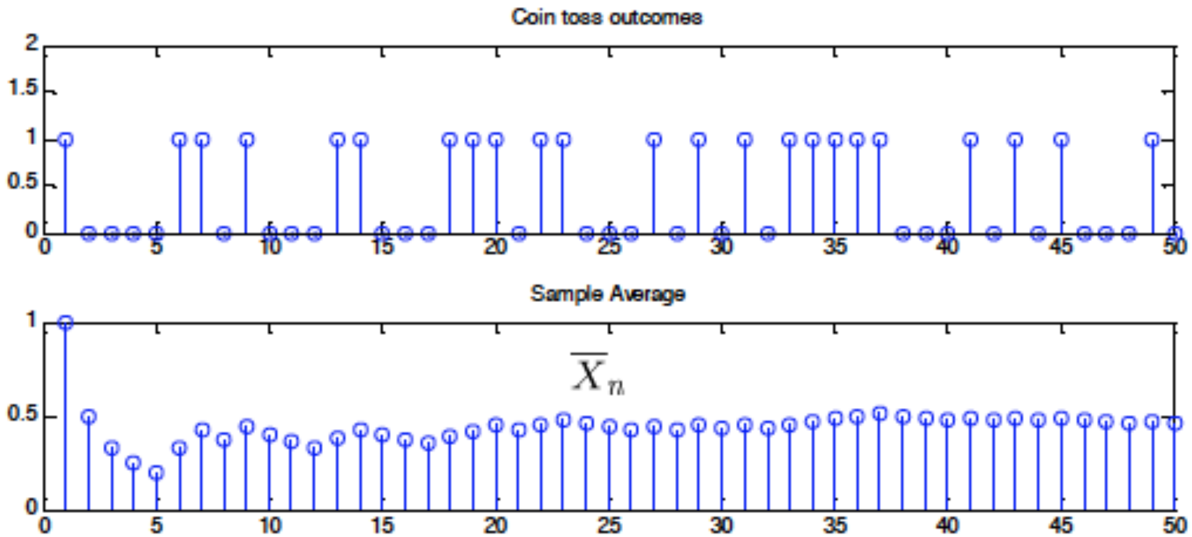
In other words, if you repeat the same experiment in a **independent** way, the average of the outcomes is going to be very close to the mean !!!

Demonstration

Example

Take a regular coin, and start flipping it several times, taking note if it falls heads or tails (flip the coin in such a way that you cannot predict the outcome of each trial).

$$X_i = \begin{cases} 1 & \text{if } i^{th} \text{ flip was 'tails'} \\ 0 & \text{if } i^{th} \text{ flip was 'heads'} \end{cases}$$

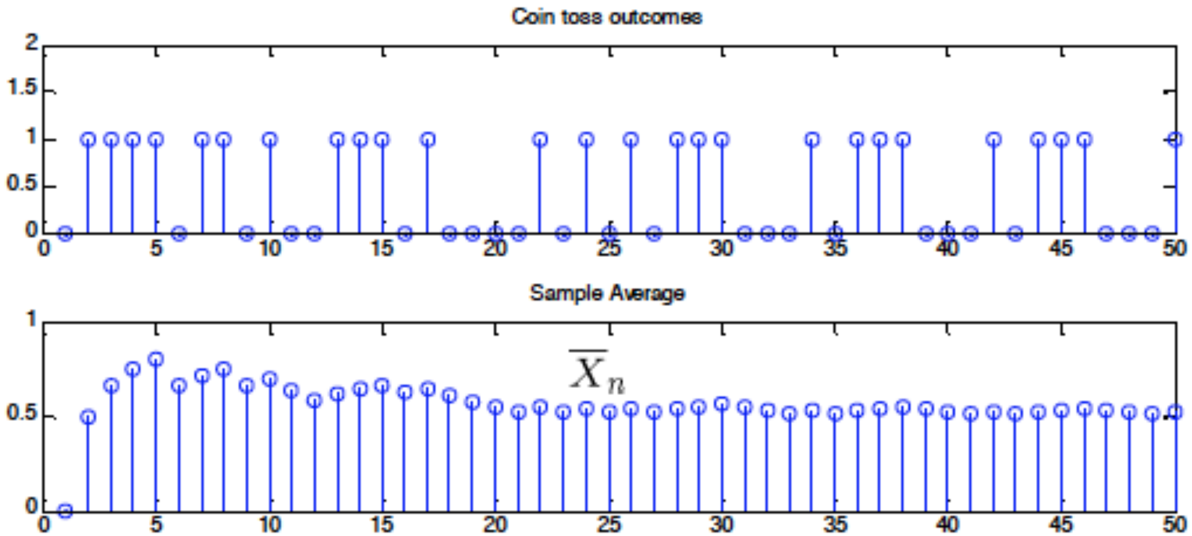


You see that the average is approaching very quickly the expected value, which is $\frac{1}{2}$ in this case!!!

Example

Take a regular coin, and start flipping it several times, taking note if it falls heads or tails (flip the coin in such a way that you cannot predict the outcome of each trial).

$$X_i = \begin{cases} 1 & \text{if } i^{th} \text{ flip was 'tails'} \\ 0 & \text{if } i^{th} \text{ flip was 'heads'} \end{cases}$$



You see that the average is approaching very quickly the expected value, which is $\frac{1}{2}$ in this case!!!

Using Data to Estimate Probabilities

The Law of Large Numbers is what allows us to use data to estimate probabilities. Suppose you stand in the bridge connecting the HG with the Auditorium, and count the number of female and male individuals you see crossing the bridge for a period of 2 hours.

$$x_i = \begin{cases} 1 & \text{if student is a female} \\ 0 & \text{if student is a male} \end{cases}$$

Say you counted 283 different individuals, and 27 were females.

Let Y be a random variable representing be the gender of a randomly choosen individual inside TU/e. Then it plausible that

$$P(Y = 1) \approx 27/283 = 0.0954 \ .$$

So provided you can make the assumption that the individuals crossing the bridge are a representative independent and identically distributed sample of Y we can use the LLN to estimate it's distribution...

The Central Limit Theorem

When discussing the normal distribution we informally stated that the sum of independent random variables resembles a normal random variable. Let's try to make this a bit more concrete

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed (i.i.d.) random variables, with mean $\mu_X = \mathbb{E}[X_i]$, and variance $\mathsf{V}(X_i) = \sigma_X^2$. Define

$$S_n = X_1 + X_2 + \ldots + X_n = \sum_{i=1}^{n} X_i \ .$$

Note that

$$\mathbb{E}[S] = \sum_{i=1}^{n} \mathbb{E}[X_i] = n\mu_X \quad \text{and} \quad \mathsf{V}(S) = \sum_{i=1}^{n} \mathsf{V}(X_i) = n\sigma_X^2 \ .$$

Then we have that the distribution of S is approximately normal, with mean $n\mu_X$ and variance $n\sigma_X^2$, that is

$$S \approx \mathcal{N}(n\mu_X, n\sigma_X^2) \ .$$

The Central Limit Theorem

The previous result was perhaps a bit sloppy (and technically not sound). More formally, we have the following important result

Theorem: The Central Limit Theorem (CLT)

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed (i.i.d.) random variables, with mean $\mu = \mathbb{E}[X_i]$, and variance $\mathsf{V}(X_i) = \sigma^2$. Define the random variable

$$Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma} \ .$$

Then, as n grows, this random variable resembles more and more a standard normal random variable.

In particular, for large n we have, for any arbitrary set B

$$P(Z_n \in B) \approx P(Z \in B) \ ,$$

where Z is a standard normal random variable.

That’s all