today: finish prob
• Frequentist vs. Bayesian

**Binomial distribution:**

- **Model:** $n$ independent coin flips
- **Random Variable:** $X_i \sim \text{Bern}(\theta)$, i.i.d.
- **Sum:** $X = \sum_{i=1}^{n} X_i$
- **Distribution:** $X \sim \text{Bin}(n, \theta)$

**PMF:**
$$p(x; \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

**Mean:**
$$\text{E}[X] = \sum_{x=0}^{n} x \cdot p(x; \theta) = n \theta$$

**Variance:**
$$\text{Var}(X) = \sum_{x=0}^{n} (x - \text{E}[X])^2 \cdot p(x; \theta) = n \theta (1-\theta)$$

**Other distributions:**
- **Poisson** ($\lambda$): $X \sim \text{Poi}(\lambda)$
- **Gaussian in 1D**: $X \sim \mathcal{N}(\mu, \sigma^2)$
- **Gamma** ($\alpha, \beta$): $X \sim \text{Gamma}(\alpha, \beta)$
Statistical concepts

- model
- data
- statistics
- prob. theory
- well-defined
- inverse problem
- Bayes
- subjective
- frequentist
- traditional
- limiting frequency
- probability
- beliefs
- subjective
- rational
- utility
- decision theory
- gambles

Example: model n independent coin flips
- probability \( \rightarrow \) prob. k heads in a row
- statistics: I want observed k heads, what is \( \theta \)? n-k tails

Semantic of prob: meaning of a prob?

a) (traditional) Frequentist semantic
- \( P_x^n = \lim \frac{X}{n} \) represents the limiting frequency of observing \( X = x \)
- if I could repeat \( n \) times of iid experiments

b) Bayesian (subjective) semantic
- \( P_x = \alpha \) encode an agent's belief that \( X = x \)
- laws of prob. characterize a "rational" way to combine "beliefs" and "evidence"
  - [observed]

Operationally,...
Bayesian approach: (very simple philosophically)

- treat all uncertain quantities as R.V.

i.e., encode all knowledge about the system ("beliefs")
as a "prior" in probabilistic models
and then use Bayes' rule to get
updated beliefs and answers.

Justification for frequentist sensors:

- for discrete R.V. $X$, suppose $P_i^\Theta x=x_i^\Theta = \Theta$
  $\Rightarrow P_i^\Theta x=x_i^\Theta = 1-\Theta$

  $\mathbb{D}^\Theta \equiv \{x_i^\Theta = x_i\}$ \Rightarrow Bin(\Theta) R.V.

  (in order to $A(u) = 1$ if $u \in \mathbb{D}$
  $\exists_0 \text{ s.w.}$

  repeat i.i.d. experiments i.e. $\Theta \overset{i.i.d.}{\sim} \text{Bin}(\Theta)$

  by L.L.N. $\frac{1}{n} \sum_{i=1}^n B_i^\Theta \overset{a.s.}{\to} \mathbb{E}[B_i^\Theta] = \Theta$

  (law of large numbers) \overset{\text{limiting frequency}}{\Rightarrow}

  by CLT $\sqrt{n} \left( (\text{Bin}(n,\Theta) - \Theta) \right) \overset{d}{\to} N(0, \frac{1}{n} \Theta)$

Coin flips - Bayesain approach

biased coin flip

unknown \Rightarrow model it as R.V.

we believe $X \sim \text{Bin}(n, \Theta)$ \Rightarrow need a $p(\Theta)$ "prior distribution"

$\Omega = [0, 1]$

Suppose we observe $X=x$ (result of $n$ flips)

then we can "update" our belief about $\Theta$ using Bayes rule

$p(\Theta = \Theta | X = x) = \frac{p(X = x | \Theta) p(\Theta = \Theta)}{p(x)}$

Posterior belief

prior belief

observation model

"marginal likelihood"
Example:

Suppose $p(\theta)$ is uniform on $[0, 1]$ "no specific preference"

$$p(\theta|x) \propto \theta^x (1-\theta)^{n-x} \frac{1}{B(\theta)} \quad B(\theta) = \frac{\Gamma(a) \Gamma(b)}{\Gamma(a+b)}$$

Scaling:

$$\int_0^1 \theta^x (1-\theta)^{n-x} \, d\theta = B(x+1, n-x+1)$$

Normalization constant $\int p(\theta|x) \, d\theta = 1$ 

Here $p(\theta|x)$ is called a \textbf{Bayesian distribution}:

$$B(\theta|x, y) \propto \theta^x (1-\theta)^{n-x}$$

- Uniform distribution: $B(\theta|1, 1)$
- Posterior: $B(\theta|x+1, n-x+1)$

Exercise to the reader: If we use $B(a, b)$ as prior

Posterior will be $B(x+a, n-x+b)$

Posterior $p(\theta|x=x)$ contains all the info from data $x$. Had we used Bayesian estimation

e.g. Question: What is the probability of landing (F=1) on the next step

as a frequentist $P(F=1 | \text{data}) = \hat{\theta}$ (estimation)

as a Bayesian $p(\theta=1 | x=x) = \int p(\theta=1, \theta | x=x) \, d\theta$ (product rule)
product rule
\[ P(x \mid \theta) = \int \underbrace{P(x \mid \theta) \cdot \pi(\theta \mid x \mid \theta) \, d\theta}_{= \Theta \quad \text{posterior}} \]
(by Bayes rule)
\[ = \int \pi(\theta \mid x \mid \theta) \, d\theta = \mathbb{E}[\theta \mid x \mid \theta] \quad \text{posterior mean of } \Theta \]

* a meaningful "Bayesian" estimator of \( \Theta \)
\[ \hat{\theta}_{\text{Bayes}}(x) = \mathbb{E}[\theta \mid x \mid \theta] \quad \text{posterior mean} \]

relation: \( \hat{\theta} : \text{desirable} \rightarrow \Theta \)

**Our (in example):**
\[ p(x \mid \theta) = \text{Beta}(\theta \mid \alpha = x+1, \beta = n-x+1) \]

mean of a beta
\[ \alpha + \beta \]

Thus
\[ \hat{\theta}_{\text{Bayes}}(x) = \mathbb{E}[\hat{\theta} \mid x] = \frac{x+1}{n+2} \]

Here, biased estimator \( \mathbb{E}[\hat{\theta} \mid x] \neq \Theta \)

\[ = \mathbb{E}[x+1] = \frac{nx+1}{n+2} \]

but asymptotically unbiased \( \xrightarrow{\text{asymptotic}} \Theta \)

compare (contrast with \( \hat{\theta}_{\text{MLE}}(x) = \frac{x}{n} \)
\[ \text{[unbiased } \mathbb{E}[\hat{\theta}] = \frac{x}{n} = \Theta] \]

**to summarize:**

- As a Bayesian: get a posterior + use loaves, probabilities
- In frequentist statistics:
  - consider multiple estimators
  - MLE
  - moment matching
  - Bayesian posterior mean
  - MAP
  - regularized MLE

and then analyze their statistical properties
Maximum likelihood principle

setup: given a parametric family $p(x; \theta)$ for $\theta \in \Theta$

we want to estimate/learn $\theta$ from $x$

$$\hat{\theta}_{ML}(x) = \arg\max_{\theta \in \Theta} p(x; \theta)$$

$L(\theta) \equiv L(x; \theta)$

"likelihood function" of $\theta$

$p(x; \theta)$