today: finish prob. 
- frequentist vs. Bayesian

**binomial distribution:**

Model \( n \) independent coin flips

- sum of \( n \) independent Bern(\( p \)) random variables
- let \( X_i \sim \text{Bern}(\theta) \) independent and identically distributed
  - implicitly defining \( X_1, X_2, \ldots, X_n \)
- let \( X = \sum_{i=1}^n X_i \), then we have \( X \sim \text{Bin}(n, \theta) \)

**binomial with parameters \( n, \theta \)**

\[ X = \{0, 1, \ldots, n\} \]

pmf: \[ p(x; n, \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \quad \text{for } x \in \Omega_X \]

\[ \binom{n}{x} = \frac{n!}{x!(n-x)!} \]

\[ p(x; n, \theta) = \frac{n!}{x!(n-x)!} \theta^x (1-\theta)^{n-x} \]

number of ways to choose \( x \) elements out of \( n \)

mean: \( X = \sum_{i=1}^n X_i \)

\[ \mathbb{E}[X] = \mathbb{E}[\sum_{i=1}^n X_i] = n \mathbb{E}[X_i] = n \theta \]

Similarly, \( \text{Var}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \text{Var}(X_i) = n \theta (1-\theta) \)

Other distributions:

- Poisson (\( \lambda \)) \[ \lambda_X = \{0, 1, \ldots, \infty\} \quad \text{count data} \]
  - mean \( \lambda \)
  - variance \( \lambda \)
- Gaussian in 1D \( N(\mu, \sigma^2) \) \[ \mathbb{L}X = \mathbb{R} \]
- Gamma \( \Gamma(\alpha, \beta) \) \[ \mathbb{L}X = \mathbb{R}^+ \]
  - shape \( \alpha \), rate \( \beta \)
  - mean \( \frac{\alpha}{\beta} \)
  - variance \( \frac{\alpha}{\beta^2} \)
- Other: Student, Cauchy, exponential, beta, Dirichlet, Dirichlet on 2 elements
Statistical concepts

- Model
- Data
- Statistics
- Probability

Example: Model n independent coin flips

Probability: \( n \) heads in a row

Statistics: I have observed \( k \) heads, what is \( \theta \)?

Frequentist vs. Bayesian:

1. Semantic of \( \theta \): meaning of a prob?
   a) (traditional) frequentist semantic
      \[ P(x=x) \text{ represents the limiting frequency of observing } x = x \]
      if I could repeat the # of iid experiments
   b) Bayesian (subjective) semantic
      \[ P(x=x) \text{ encodes an agent } \text{"belief"} \text{ that } x = x \]
      the laws of prob. characterize a "rational" way to combine "beliefs" and "evidence" [observations]

Bayesian approach:

- Very simple philosophically:
  - treat all uncertain quantities as \( \text{RV} \)
  - encode all knowledge about the system ("belief") as a "prior" on probabilistic models
  - and then use law of prob. (and Bayes rule) to get updated beliefs and answers

Justification for frequentist semantic: see notes from last year

Coin flip vs. bayesian approach

Biased coin flip unknown \( \Rightarrow \) model it as a \( \text{RV} \)
biased coin flip

unknown \Rightarrow model it as a r.v.

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

\( \theta \in [0,1] \)

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

then we "update" our beliefs about \( \theta \) using Bayes rule

\[
p(\theta \mid x=x) = \frac{p(x \mid \theta) p(\theta)}{p(x)}
\]

posterior belief

\( p(x) \) normalization

\( p(x \mid \theta) \) observation model

\( p(\theta) \) "prior belief"

\( p(x) \) "marginal likelihood"

\( \text{Note: } p(x \mid \theta) \rightarrow p(x \mid \theta) \quad p(\theta) \rightarrow p(\theta) \)

\( p(x \mid \theta) \) is a "mixed distribution"

\( p(\theta) \rightarrow p(\theta) \)

Example:

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

\[
p(\theta \mid x) \propto p(x \mid \theta) p(\theta)
\]

\( p(x \mid \theta) \propto \theta^x (1-\theta)^{n-x} \)

\( p(\theta) \propto 1 \) up to a constant

Scaling:

\[
\int \theta^x (1-\theta)^{n-x} \, d\theta = B(x+1, n-x+1)
\]

\( B(a, b) = \frac{\Gamma(a) \Gamma(b)}{\Gamma(a+b)} \Gamma(a-1) = \frac{\Gamma(a) \Gamma(b)}{\Gamma(a+b)} \]

\( \Gamma(a) = \int_0^1 u^{a-1} e^{-u} \, du \)

normalization constant

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

Here \( p(\theta \mid x) \) is called a "beta distribution"

\[
B(\theta \mid a, b) \triangleq \frac{\theta^{a-1} (1-\theta)^{b-1}}{\text{Gamma}(a,b)}
\]

\( \text{parameters} \)

* Uniform distribution \( B(\theta \mid 1, 1) \)

* posterior \( p(\theta \mid x+1, n-x+1) \)

Exercise for the reader: if we \( B(\alpha, \beta) \) as prior

show that posterior will be \( B(x+\alpha, n-x+\beta) \)
posterior \( p(\theta|x=x) \) contains all the info from data \( x \) that we need to answer questions about \( \theta \)

e.g. question: what is prob. of head (\( F=1 \)) on the next flip

as a frequentist

\[ P(F=1|data) = \theta \]

as a Bayesian

\[ P(F=1|x=x) = \int p(F=1, \theta | x=x) d\theta \]

\[ = \int p(F=1|\theta, x=x) p(\theta|x=x) d\theta \]

\[ = \int \theta p(\theta|x=x) d\theta = \mathbb{E}_\theta[\theta|x=x] \]

\[ = \mathbb{E}_\theta[\theta] = \theta \]

* a meaningful "Bayesian" estimator of \( \theta \)

\[ \hat{\theta}_{\text{Bayes}}(x) = \mathbb{E}[\theta|x=x] \quad (\text{posterior mean}) \]

relation: \( \hat{\theta} : \text{observation} \rightarrow \theta \)

coin coin example: \( p(x|\theta) = \text{beta}(\theta|\alpha=\theta+1, \beta=n-\theta+1) \)

mean of a beta RV. = \( \frac{\alpha}{\alpha+\beta} \)

thus

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

here, biased estimator \( \mathbb{E}_\theta[\hat{\theta}(x)] \neq \theta \)

but asymptotically unbiased \( \frac{x+1}{n+2} \to \theta \)

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

summarize:

* as a Bayesian, get a posterior + use law of probabilities

* in 'frequentist statistics',...
Maximum likelihood principle

Step: Given a parametric family $p(x; \theta)$ for $\theta \in \Theta$.

We want to estimate/learn $\theta$ from $x$.

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

The "likelihood function" $L(\theta)$

MLE example 1: Binomials

$n$ coin-flips

$X \sim \text{Bin}(n, \theta)$

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

Trick: To maximize $\log L(\theta)$ instead of $L(\theta)$

To maximize $\log p(x; \theta)$ instead of $p(x; \theta)$

Justification: $\log(.)$ is strictly increasing.

ie. $a<b \implies \log a < \log b$ for $a, b > 0$

$$\implies \arg \max_{\theta \in \Theta} \log p(x; \theta) = \arg \max_{\theta \in \Theta} p(x; \theta)$$

$$\log p(x; \theta) = \log \left( \binom{n}{x} \right) + x \log \theta + (n-x) \log (1-\theta) = \ell(\theta)$$

$$\ell'(\theta) = 0 \implies \frac{d}{d \theta} \ell(\theta) = 0$$

Look for $\theta$ s.t. $\theta_{MLE}$

Want

$\ell' = \frac{d}{d \theta} \ell(\theta) = 0$

$\theta \in \Theta$

$x(1-\theta) - \theta(n-x) = 0$

$\theta = \frac{x}{n}$
Some optimization comment:

\[ \min_{x \in \mathbb{R}} f(x) \quad \text{if } f \text{ is diff.} \]

\[ \nabla f(x) = 0 \quad \text{is a necessary condition for } x \text{ being a local min when } x \text{ is in the interior of } \mathbb{R}^n \]

Also need to check

\[ \nabla^2 f(x) > 0 \]

Hessian \( \nabla^2 f(x) > 0 \)

Hessian is positive definite with \( \text{Hu} > 0 \)

\[ \text{only local result in general} \]

- If \( \nabla^2 f(x) > 0 \) \( \forall x \in \mathbb{R}^n \), \( f(x) \) is said "convex".
- In this case, \( \nabla f(x) = 0 \) \( \Rightarrow \) sufficient for \( x \) to be a global min.

- Otherwise, for smooth \( f(x) \), lack of zero gradients and boundary points gives enough information to find global optima.