# 18.650 – Fundamentals of Statistics

## 6. Bayesian Statistics

# Goals

So far, we have followed the *frequentist* approach (cf. meaning of a confidence interval).

An alternative is the **Bayesian approach**.

New concepts will come into play:

- prior and posterior distributions
- Bayes' formula
- Priors: improper, non informative
- Bayesian estimation: posterior mean, Maximum a posteriori (MAP)
- Bayesian confidence region

In a sense, Bayesian inference amounts to having a likelihood function $L_n(\theta)$ that is weighted by prior knowledge on what $\theta$ might be. This is useful in many applications.

# The frequentist approach

- Assume a statistical model $(E, \{\mathbb{P}_\theta\}_{\theta \in \Theta})$.
- We assumed that the data $X_1, \ldots, X_n$ was drawn i.i.d from $\mathbb{P}_{\theta^*}$ for some unknown **fixed** $\theta^*$.
- When we used the MLE for example, we looked at all possible $\theta \in \Theta$.
- Before seeing the data we did not prefer a choice of $\theta \in \Theta$ over another.

# The Bayesian approach

- In many practical contexts, we have a **prior belief** about $\theta^*$
- Using the data, we want to update that belief and transform it into a **posterior belief**.

# The kiss example

- Let $p$ be the proportion of couples that turn their head to the right
- Let $X_1, \ldots, X_n \overset{i.i.d}{\sim}$ Ber($p$).
- In the frequentist approach, we estimated $p$ (using the MLE), we constructed some confidence interval for $p$, we did hypothesis testing (e.g., $H_0 : p = .5$ v.s. $H_1 : p \neq .5$).
- Before analyzing the data, we may believe that $p$ is likely to be close to $1/2$.
- The Bayesian approach is a tool to update our prior belief using the data.

# The kiss example

- Our prior belief about $p$ can be quantified:

- E.g., we are 90% sure that $p$ is between $.4$ and $.6$, 95% that it is between $.3$ and $.8$, etc...

- Hence, we can model our prior belief using a distribution for $p$, *as if* $p$ was random.

- In reality, the true parameter is not random ! However, the Bayesian approach is a way of modeling our belief about the parameter by doing **as if** it was random.

- E.g., $p \sim \text{Beta}(a, b)$ (*Beta distribution*). It has pdf

$$f(x) = \frac{1}{K} x^{a-1}(1-x)^{b-1} \mathbb{1}(x \in [0,1]), \quad K = \int_0^1 t^{a-1}(1-t)^{b-1} \, \mathrm{d}t$$

- This distribution is called the **prior distribution**

# The kiss example

- In our statistical experiment, $X_1, \dots, X_n$ are assumed to be i.i.d. Bernoulli r.v. with parameter $p$ **conditionally on** $p$.

- After observing the available sample $X_1, \dots, X_n$, we can update our belief about $p$ by taking its distribution conditionally on the data.

- The distribution of $p$ conditionally on the data is called the *posterior distribution*.

- Here, the posterior distribution is

$$\mathsf{Beta}\Big(a + \sum_{i=1}^{n} X_i, b + n - \sum_{i=1}^{n} X_i\Big)$$

# Clinical trials

Let us revisit our clinical trial example

- Pharmaceutical companies use hypothesis testing to test if a new drug is efficient.
- To do so, they administer a drug to a group of patients (test group) and a placebo to another group (control group).
- We consider testing a drug that is supposed to lower LDL (low-density lipoprotein), a.k.a "bad cholesterol" among patients with a high level of LDL (above 200 mg/dL)

# Clinical trials

- Let $\Delta_{\mathrm{d}} > 0$ denote the expected decrease of LDL level (in mg/dL) for a patient that has used the drug.
- Let $\Delta_{\mathrm{c}} > 0$ denote the expected decrease of LDL level (in mg/dL) for a patient that has used the placebo.

Quantity of interest: $\theta := \Delta_d - \Delta_c$.

In practice we have a prior belief on $\theta$. For example,

- $\theta \sim \mathsf{Unif}([100, 200])$
- $\theta \sim \mathsf{Exp}(1/100)$
- $\theta \sim \mathcal{N}(100, 300)$,
- $\dots$

# Prior and posterior

- Consider a probability distribution on a parameter space $\Theta$ with some pdf $\pi(\cdot)$: the *prior distribution*.
- Let $X_1, \ldots, X_n$ be a sample of $n$ random variables.
- Denote by $L_n(\cdot|\theta)$ the joint pdf of $X_1, \ldots, X_n$ conditionally on $\theta$, where $\theta \sim \pi$.
- **Remark:** $L_n(X_1, \ldots, X_n|\theta)$ is the likelihood used in the frequentist approach.
- The conditional distribution of $\theta$ given $X_1, \ldots, X_n$ is called the *posterior distribution*. Denote by $\pi(\cdot|X_1, \ldots, X_n)$ its pdf.

# Bayes' formula

- Bayes' formula states that:

$$\pi(\theta|X_1,\ldots,X_n) \propto \pi(\theta)L_n(X_1,\ldots,X_n|\theta), \quad \forall\theta\in\Theta.$$

- The constant does not depend on $\theta$:

$$\pi(\theta|X_1,\ldots,X_n) = \frac{\pi(\theta)L_n(X_1,\ldots,X_n|\theta)}{\int_\Theta L_n(X_1,\ldots,X_n|t)\pi(t)\,\mathrm{d}t}, \quad \forall\theta\in\Theta.$$

# Bernoulli experiment with a Beta prior

In the Kiss example:

- $p \sim \mathsf{Beta}(a, a)$:

$$\pi(p) \propto p^{a-1}(1-p)^{a-1}, p \in (0,1)$$

- Given $p$, $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathsf{Ber}(p)$, so

$$L_n(X_1, \ldots, X_n | p) = p^{\sum_{i=1}^n X_i}(1-p)^{n-\sum_{i=1}^n X_i}.$$

- Hence,

$$\pi(p|X_1, \ldots, X_n) \propto p^{a-1+\sum_{i=1}^n X_i}(1-p)^{a-1+n-\sum_{i=1}^n X_i}.$$

- The posterior distribution is

$$\mathsf{Beta}\left(a + \sum_{i=1}^n X_i, a + n - \sum_{i=1}^n X_i\right).$$

# Non informative priors

- We can still use a Bayesian approach if we have no prior information about the parameter. How to pick prior $\pi$?
- Good candidate: $\pi(\theta) \propto 1$, i.e., constant pdf on $\Theta$.
- If $\Theta$ is bounded, this is the uniform prior on $\Theta$.
- If $\Theta$ is unbounded, this does not define a proper pdf on $\Theta$ !
- An *improper prior* on $\Theta$ is a measurable, nonnegative function $\pi(\cdot)$ defined on $\Theta$ that is not integrable.
- In general, one can still define a posterior distribution using an improper prior, using Bayes' formula.

# Examples

- If $p \sim \mathsf{Unif}(0,1)$ and given $p$, $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathsf{Ber}(p)$ :

$$\pi(p|X_1, \ldots, X_n) \propto p^{\sum_{i=1}^n X_i}(1-p)^{n-\sum_{i=1}^n X_i}$$

  i.e., the posterior distribution is

$$\mathsf{Beta}\left(1+\sum_{i=1}^n X_i, 1+n-\sum_{i=1}^n X_i\right).$$

- If $\pi(\theta) = 1, \forall \theta \in \mathbb{R}$ and given $X_1, \ldots, X_n|\theta \overset{i.i.d.}{\sim} \mathcal{N}(\theta, 1)$:

$$\pi(\theta|X_1, \ldots, X_n) \propto \exp\left(-\frac{1}{2}\sum_{i=1}^n (X_i - \theta)^2\right)$$

  i.e., the posterior distribution is

$$\mathcal{N}\left(\bar{X}_n, \frac{1}{n}\right).$$

# Bayesian confidence regions

- For $\alpha \in (0,1)$, a Bayesian confidence region with level $\alpha$ is a random subset $\mathcal{R}$ of the parameter space $\Theta$, which depends on the sample $X_1, \ldots, X_n$, such that:

$$\mathbb{P}[\theta \in \mathcal{R} | X_1, \ldots, X_n] = 1 - \alpha.$$

- Note that $\mathcal{R}$ depends on the prior $\pi(\cdot)$.

- "Bayesian confidence region" and "confidence interval" are two **distinct** notions.

# Bayesian estimation

- The Bayesian framework can also be used to estimate the true underlying parameter (hence, in a frequentist approach).
- In this case, the prior distribution does not reflect a prior belief: It is just an artificial tool used in order to define a new class of estimators.
- **Back to the frequentist approach:** The sample $X_1, \ldots, X_n$ is associated with a statistical model $(E, (\mathbb{P}_\theta)_{\theta \in \Theta})$.
- Define a prior (that can be improper) with pdf $\pi$ on the parameter space $\Theta$.
- Compute the posterior pdf $\pi(\cdot | X_1, \ldots, X_n)$ associated with $\pi$.

# Bayesian estimation

- *Bayes estimator:*

$$\hat{\theta}^{(\pi)} = \int_{\Theta} \theta \, \mathrm{d}\pi(\theta | X_1, \ldots, X_n)$$

  This is the *posterior mean*.

- The Bayesian estimator depends on the choice of the prior distribution $\pi$ (hence the superscript $\pi$).

- Another popular choice is the point that maximizes the posterior distribution, provided it is unique. It is called the MAP (maximum a posteriori):

$$\hat{\theta}^{\mathrm{MAP}} = \underset{\theta \in \Theta}{\operatorname{argmax}} \, \pi(\theta | X_1, \ldots, X_n)$$

# Bayesian estimation

- In the previous examples:
  - Kiss example with prior Beta$(a, a)$ $(a > 0)$:

    $$\hat{p}^{(\pi)} = \frac{a + \sum_{i=1}^{n} X_i}{2a + n} = \frac{a/n + \bar{X}_n}{2a/n + 1}.$$

    In particular, for $a = 1/2$ (Jeffreys prior),

    $$\hat{p}^{(\pi_J)} = \frac{1/(2n) + \bar{X}_n}{1/n + 1}.$$

  - Gaussian example with improper prior $\pi(\theta) \propto 1$: $\hat{\theta}^{(\pi_J)} = \bar{X}_n$.
- In each of these examples, the Bayes estimator is consistent and asymptotically normal.
- In general, the asymptotic properties of the Bayes estimator do not depend on the choice of the prior.