Unit #1: Introduction to Bayesian Inference

Bayesian Statistical Methods, Chapter 1

Photo by Glenn Asakawa/University of Colorado

Photo by Jana Knorr on Unsplash

A **stopping rule** is a rule that tells a researcher when to conclude data collection.

Example: Thom is collecting data on coffee sales at a local coffee shop. The purpose of his analysis is to infer the how coffee sales vary by demographic (e.g., do younger or older customers spend more money at the coffee shop?). He decides to observe coffee sales *until he's finished with his cup of coffee*.

Photo by Jana Knorr on Unsplash

Imagine health researchers are investigating the prevalence of a rare genetic condition, denoted as G. If the percentage of individuals carrying the G gene in the population, p, is greater than 0.5, certain recommendations might be given to insurance companies to cover further screenings for G. The researchers set the significance level at $\alpha = 0.1$.

They collected the following binary data, where $1 =$ "does not carry G" and $0 =$ "carries G":

$$\mathbf{x} = (0,0,0,1,0,1,0,0,0,1)$$

What was the stopping rule?

Photo by Jana Knorr on Unsplash

Activity: Should “stopping rules” matter? Do they matter in frequentist inference?

Photo by Jan Huber on Unsplash

Probability and statistics review:

- Basic theorems of probability (including Bayes' theorem for events!).
- Univariate distributions/random variables (both discrete and continuous).
- **Multivariate distributions.**
- Random samples, distributions of estimators/test statistics, the central limit theorem.
- Estimation theory (maximum likelihood estimation, properties of estimators)
- Confidence intervals.
- Hypothesis testing.

Photo by Jan Huber on Unsplash

An n-dimensional vector $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ as a **multivariate normal distribution** with mean vector $E(\mathbf{Y}) = \boldsymbol{\mu}$ and variance covariance matrix $Var(\mathbf{Y}) = \Sigma$ if, for any constant (i.e., *non-random*) vector $\mathbf{a} = (a_1, a_2, \ldots, a_n)^T$, every linear combination $X = \mathbf{a}^T\mathbf{Y} = a_1Y_1 + a_2Y_2 + \ldots + a_nY_n$ has a univariate normal distribution with mean $\mu_X = \mathbf{a}^T\boldsymbol{\mu}$ and variance $\sigma_X^2 = \mathbf{a}^T\Sigma\mathbf{a}$.

Photo by Jan Huber on Unsplash

An n-dimensional multivariate normal vector $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ can be written in the following notation: $\mathbf{Y} \sim N\left(\boldsymbol{\mu}, \Sigma\right)$, where

$$\boldsymbol{\mu} = (\underbrace{\mu_1}_{E(Y_1)}, \ldots, \underbrace{\mu_n}_{E(Y_n)})^T$$

is a vector of means and variance-covariance matrix given by

$$\Sigma = \begin{pmatrix} \sigma_{1,1} & \sigma_{1,2} & \cdots & \sigma_{1,n} \\ \sigma_{2,1} & \sigma_{2,2} & \cdots & \sigma_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n,1} & \sigma_{n,2} & \cdots & \sigma_{n,n} \end{pmatrix}.$$

The entries of Σ are:

$\sigma_{i,j} = Cov(Y_i, Y_j)$ for $i, j = 1,\ldots,n$.

Photo by Jan Huber on Unsplash

Example: Let $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ be an n -dimensional multivariate normal vector: $\mathbf{Y} \sim N\left(\boldsymbol{\mu}, \Sigma\right)$, where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$ is a vector of means and variance-covariance matrix given by Σ.

Photo by Jan Huber on Unsplash

Conditional distributions:

The conditional distribution of X_1 given that the random variable X_2 is fixed at x_2 is given by:

$$f_{X_1 \mid X_2}\left(x_1 \,\middle|\, X_2 = x_2\right) = \frac{f(x_1, x_2)}{f(x_2)}, \quad f(x_2) > 0.$$

Photo by Jan Huber on Unsplash

Conditional from multivariate normal:

Let $(Y_1, Y_2) \sim N(\boldsymbol{\mu}, \Sigma)$.

Let $\rho = \dfrac{\sigma_{1,2}}{\sqrt{\sigma_{1,1}\sigma_{2,2}}} = \dfrac{Cov(Y_1, Y_2)}{\sqrt{Var(Y_1)Var(Y_2)}}$. Then the distribution $Y_2 \mid Y_1 = y_1$ is given by

$$Y_2 \mid Y_1 \sim N\left(\mu_2 + \rho\sqrt{\frac{\sigma_{22}}{\sigma_{11}}}\left(y_1 - \mu_1\right), (1-\rho^2)\sigma_{22}\right)$$

Photo by Jan Huber on Unsplash

Let (Y_1, Y_2) be bivariate normal with $E(Y_1) = E(Y_2) = 0$, $Var(Y_1) = Var(Y_2) = 1$, and $Cov(Y_1, Y_2) = \rho$.

1. What is the distribution of Y_1?

2. Derive the conditional distribution of $Y_1 \mid Y_2$.

Photo by Jan Huber on Unsplash

How is probability theory used in statistical inference?

Photo by Jan Huber on Unsplash

“Should probability enter [statistics] to capture degrees of belief about claims? To measure variability? Or to ensure we won’t reach mistaken interpretations of data too often in the long run of experience? Modern statistical methods grew out of attempts to systematize doing all of these.”

—Deborah Mayo, *Statistical Inference as Severe Testing*

Photo by Jan Huber on Unsplash

“The field [of statistics] has been marked by disagreements between competing … [groups] of frequentists and Bayesians that have been so contentious – likened in some quarters to religious and political debates – that everyone wants to believe we are long past them…The truth is, long-standing battles still simmer below the surface in questions about scientific trustworthiness and the relationships between Big Data-driven models and theory.”

—Deborah Mayo, *Statistical Inference as Severe Testing*

Photo by Jan Huber on Unsplash

What is frequentist statistical inference?

Photo by Jan Huber on Unsplash

“By and large, classical statistical procedures share the feature that they only rely on probability assignments over sample spaces…an important motivation for this is that those probabilities can be interpreted as frequencies.” —Jan-Willem Romeijn, https://stanford.io/3JJD6rS

Photo by Jan Huber on Unsplash

A frequentist statistical model

$$\left(\mathbf{X}\,,\, f(\mathbf{x};\theta)\right)$$

Random variables that generate the data.

The joint probability distribution over the random variables, for a fixed set of parameters, $\boldsymbol{\theta}$.

"The *statistical model* is chosen to represent a description of a chance mechanism that accounts for the systematic information (the chance regularities) in the data." —Aris Spanos

Photo by Jan Huber on Unsplash

Frequentist methods in statistics:

1. *Hypothesis testing*. z-test, t-test, chi-squared test, F-test.

2. *Point estimation*. Maximum likelihood estimation (in one context, least squares), method of moments.

3. *Interval estimation*. Confidence intervals.

From: https://bit.ly/3gTjlBg

What is a Bayesian statistics?

From: https://bit.ly/3gTjlBg

"By and large, classical statistical procedures share the feature that they only rely on probability assignments over sample spaces…an important motivation for this is that those probabilities can be interpreted as frequencies." — Jan-Willem Romeijn, https://stanford.io/3JJD6rS

"The defining characteristic of Bayesian statistics is that it considers probability distributions over statistical hypotheses as well as over data."—Jan-Willem Romeijn, https://stanford.io/3qIarKQ

From: https://bit.ly/3gTjlBg

A Bayesian statistical model

$$\left(\mathbf{X}\,,\, f(\mathbf{x}, \theta)\right)$$

Random variables that generate the data

Joint probability distribution over the data **and** parameters

From: https://bit.ly/3gTjlBg

Bayesian methods in statistics:

1. *Posterior distributions for a parameter (or set of parameters).* Computed using Bayes' theorem.

2. *Point estimation*. Uses the posterior distribution to find a single value to estimate an unknown parameter.

3. *Interval estimation*. Uses the posterior distribution to construct a range of values that, with some probability, contains the unknown parameter.

4. *Bayes' factors*. Provide a method for testing hypotheses.

Comparisons of frequentist and Bayesian statistics

Frequentist:

- Parameters are unknown and fixed.
- Probability theory describes the processes that generate data, and is *not* applied to parameters/ hypotheses.
- The chance processes that generate data are used to estimate parameters or make inferences about hypotheses related to those parameters.
- We assess the *frequentist properties* of estimators and test statistics. For example, is the estimator unbiased? Consistent? Efficient?
- These properties have to do w/*frequency,* i.e., repeated sampling.
  - Example: if I chose a sample of size n over and over from this population, would my estimator be correct, on average?

Bayesian:

- Parameters are unknown and fixed, but we *model* them as random variables.
  - This model is usually thought of as representing our *degree of belief* in the value of the parameter.
  - We might think (before sampling): I don't know the (fixed) mean height of all CU students, but I believe that it's definitely between 5 and 6 ft. (Prior Belief).
- We use Bayes' theorem to combine our prior belief about a parameter with evidence from a sample.
- The result is a new probability distribution over the parameter. It represents our belief in the value of the parameter after accounting for evidence (the sample). The mean or mode might serve as an estimator..
- We can assess the frequentist performance (e.g., unbiasedness) of a Bayesian estimator.

From: https://bit.ly/3gTjlBg

In the frequentist approach to estimation, hypotheses, represented statistically by model parameters, θ, were thought to be unknown and fixed (i.e., not random).

In the Bayesian approach, parameters are considered to be a quantities whose variation can be described by a (subjective) probability distribution (called a prior distribution). A sample is then taken from the population and the prior is updated—using Bayes' theorem—to obtain a posterior distribution.

From: https://bit.ly/3gTjlBg

$$\left(\mathbf{X}\,,\, f(\mathbf{x}, \theta)\right)$$

Random variables that generate the data

Joint probability distribution over the data **and** parameters

From: https://bit.ly/3gTjlBg

$$\left(\mathbf{X}\,,\, f(\mathbf{x}, \theta)\right)$$

Random variables that generate the data

Joint probability distribution over the data **and** parameters

$f(\mathbf{x}, \boldsymbol{\theta})$ is the joint probability of the data, $\mathbf{X}$, and the parameters, $\boldsymbol{\theta}$.

Denote the marginal distribution of $\boldsymbol{\theta}$ as $\pi(\boldsymbol{\theta})$, the marginal distribution of $\mathbf{X}$ as $f(\mathbf{x})$, and the conditional distributions as $\pi(\boldsymbol{\theta} \mid \mathbf{x})$ and $f(\mathbf{x} \mid \boldsymbol{\theta})$.

Then:

The marginal distribution of $\boldsymbol{\theta}$, $\pi(\boldsymbol{\theta})$, is called the **prior distribution**.

The conditional distribution of $\boldsymbol{\theta}$ given the sample $\mathbf{x}$, $\pi(\boldsymbol{\theta} \mid \mathbf{x})$, is called the **posterior distribution.**

The conditional distribution of $\mathbf{x}$ given the parameters $\boldsymbol{\theta}$, $f(\mathbf{x} \mid \boldsymbol{\theta})$, is called the **likelihood function**.

Notes on the prior distribution:

1. The prior distribution is meant to reflect our knowledge/degree of belief in the value of $\boldsymbol{\theta}$ before any new data are observed.

2. It is often the case that we have prior knowledge, an expert opinion, etc. $\pi(\boldsymbol{\theta})$ is the formal way to take that knowledge into account.

3. There is a *ton* to say about what to do in the absence of prior information!

    1. "Uninformative" priors?

    2. "Objective" priors?

Notes on the likelihood function:

1. The Bayesian likelihood function is (almost) conceptually the same as the frequentist likelihood function (e.g., in MLE).

2. The likelihood gives the distribution of the sample given the parameter(s), but interpreted as a function of the parameters.

Notes on the posterior distribution:

1. The posterior distribution describes our uncertainty about the parameter(s) given the data (and our prior knowledge).

2. It is used to make inferential statements about $\boldsymbol{\theta}$, which is still considered a random quantity. In Bayesian inference, an inference about $\boldsymbol{\theta}$ is only valid if it is made based on the posterior distribution.

3. We can construct estimators/estimates of $\boldsymbol{\theta}$ using the posterior, e.g., the mean or mode of the posterior can be used as a point estimate of $\boldsymbol{\theta}$.

4. We can use the posterior to construct interval estimates of $\boldsymbol{\theta}$.

From: https://bit.ly/3gTjlBg

A note on the denominator of Bayes' theorem:

1. It is often very difficult to compute $f(\mathbf{x})$ directly, which is why we attempt to compute $f(\mathbf{x}) = \int f(\mathbf{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$ (or $f(\mathbf{x}) = \sum_{\boldsymbol{\theta}} f(\mathbf{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})$ when $\boldsymbol{\theta}$ is discrete).

2. $f(\mathbf{x}) = \int f(\mathbf{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$ itself can be difficult to compute, especially when $\boldsymbol{\theta}$ is high-dimensional (e.g., imagine a complicated statistical learning model with hundreds or thousands of parameters…that's a tough integral!).

3. We will study techniques for avoiding the computation of this integral.

From: https://bit.ly/3gTjlBg

Activity: Open “Unit #1: Introduction to Bayesian Inference” from Canvas and work in groups of 2-3 on:

1. Problem #1(f) (continuing our discussion of stopping rules).

2. Problem #2 (a discrete example of Bayesian inference, and a connection to maximum likelihood).

Suppose that we have a posterior distribution, $\pi(\boldsymbol{\theta} \mid \mathbf{x})$.

How can we summarize it? Which summaries will constitute reasonable estimators of $\boldsymbol{\theta}$?

Suppose that we have a posterior distribution, $\pi(\boldsymbol{\theta} \mid \mathbf{x})$.

1. The **maximum a posteriori (MAP) estimator**:

2. The **mean/expectation of the posterior distribution**:

Suppose that we have a posterior distribution, $\pi(\theta \mid \mathbf{x})$, and we want an interval that we are $95\,\%$ certain contains the true parameter, θ, given the data.

Let $Q(\theta)$ be the cdf of $\boldsymbol{\theta} \mid \mathbf{x}$ and $Q^{-1}(x)$ be the inverse cdf of $\boldsymbol{\theta} \mid \mathbf{x}$. Then we can write a $95\,\%$ **credible interval estimator** as:

Photo by Ryan Stone on Unsplash

The goal of **explanatory modeling** is to makes use of statistical models for testing and providing causal explanations.

Photo by Ryan Stone on Unsplash

Explanatory models often:

1. Have parameters with a **physical meaning** (e.g., Hooke's Law).

2. Arise in the context of an **experiment design**.

3. Require additional assumptions if the data come from an **observational study**.

Photo by Ryan Stone on Unsplash

A **prediction** is an empirical consequence of a theory that has not yet been verified at the time the theory was constructed.

A **predictive model** is a statistical model used to provide a value (or range of values) of the response based on values of the predictors that were not used to train the model.

Photo by Ryan Stone on Unsplash

Like frequentist methodology, Bayesian methodology can be employed for explanation or prediction.

Let $\left(\mathbf{X}, f(\mathbf{x};\theta)\right)$ be a *frequentist* statistical model, and let $\mathbf{X}^*$ be a future observation that we would like to predict. Assume $\mathbf{X}^*$ is independent of $\mathbf{X}$ (e.g., we don't have a time series).

In frequentist inference, we would use an estimate of θ, denoted θ, and the standard error of θ, to derive point and interval estimates for $\mathbf{X}^*$.

Photo by Ryan Stone on Unsplash

Let $\left(\mathbf{X}\,,\, f(\mathbf{x}, \theta)\right)$ be our *Bayesian* statistical model, and let $\mathbf{X}^*$ be a future observation that we would like to predict. Assume $\mathbf{X}^*$ is independent of $\mathbf{X}$ (e.g., we don't have a time series).

Bayesian will derive a *distribution* of possible values for $\mathbf{X}^*$. This distribution will take into account the variability in the data generating process *and* variability in θ.

Photo by Ryan Stone on Unsplash

The **posterior predictive distribution**, $f(y^* \mid \mathbf{y})$, is the distribution of future observations, $\mathbf{X}^*$, given the observed data.

Derivation:

Photo by Ryan Stone on Unsplash

Example in R!