Table of contents

Topics covered in section {#topics}

- Bayesian and frequentist views
  - Likelihood functions
- Estimands, estimators, and estimates
  - Bias and standard error
- Maximum likelihood estimation
- Method of moments

Stat 110 material {#stat110}

- This is for reference and will *not* be covered in section
- The following list is non-exhaustive
- Bayes' rule $P(A \mid B) = \frac{P(B|A)P(A)}{P(B)}$
  - Odds form: $\frac{P(A|B)}{P(A^c|B)} = \frac{P(B|A)}{P(B|A^c)} \frac{P(A)}{P(A^c)}$
- Linearity of expectation
  - Applies even when variables are dependent
- Fundamental bridge
  - $\mathbb{E}[\mathbb{1}(X = a)] = P(X = a)$

- Variance of linear combinations
    - Variance when 1st and 2nd moments exist: $\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$.
    - Variance of linear combinations:
      $\mathrm{Var}\left(a + \sum_{i=1}^n b_i X_i\right) = \sum_{i=1}^n b_i^2 \mathrm{Var}(X_i) + \sum_{1 \leq i < j \leq n} 2 b_i b_j \mathrm{Cov}(X_i, X_j)$
- Sample variance
    - Division by $n - 1$, not n
- Law of large numbers (LLN)
    - Let $X_1, \ldots, X_n$ (i.i.d.) have finite mean μ and variance σ^2, as $n \to \infty$
    - Weak LLN: $\bar{X}_n \xrightarrow{p} \mu$ (convergence in probability)
        - $P(\mid \bar{X}_n - \mu \mid > \epsilon) \to 0$
    - Strong LLN: convergence almost surely
        - $P(\lim_{n \to \infty} \bar{X}_n \to \mu) = 1$
- Central limit theorem
    - $\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} \mathcal{N}(0, 1)$.
- Deriving joint, marginal, and conditional distributions
    - Chapter 7 in Stat 110

Stat 111 basics {#basics}

Key vocabulary

- Estimand: Object that we are trying to predict. It is the quantity of interest and is usually unobservable.
- Data: Observed values. Often generated probabilistically.
- Model: (Probabilistic) assumptions about the distribution of the data.
- Estimators: Statistics that approximate the estimand. Estimators are functions of the data.
- Estimate: The estimator evaluated at observed values of the data.

Benchmarking: Bias, variance, standard error {#benchmarks}

- **Bias**: $\mathrm{Bias}_\theta(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$.
    - **Unbiased** if the above is zero
    - The average difference of an estimator from the estimand
    - Unbiased $\neq$ a good estimator
        - Can you think of examples?
    - Can we just add/substract the bias?
        - Usually depends on θ

- **Standard error**: $\mathrm{SE}(\hat{\theta}) = \sqrt{\mathrm{Var}(\hat{\theta})}$
    - Same as standard deviation
    - Terminology diffrentiates it from single observation
- **Consistency**: An estimator $\hat{\theta}$ is consistent if as $n \to \infty$, we have that $\hat{\theta} \to \theta$ in probability
- Show consistency by applying LLN and/or showing that the MSE goes to zero.
- **Bias-variance tradeoff**: decreasing the bias of an estimator comes at the cost of increasing variance

Bayesian and frequentist views {#bayesfreq}

- Let θ be our estimand
- The frequentist approach focuses on procedures that work in the long run
    - Redrawing samples
- Bayesian focuses on data at hand
    - Specifies a prior and calculates a posterior distribution of θ

Frequentist

- Treat θ as fixed
- Likelihood function assess plausibility of various possible values of θ
- Uses likelihood function for maximum likelihood estimation (MLE) and likelihood ratios

Bayesian

- Let θ be our estimand
- Bayesian reasoning adopts a prior distribution on θ to characterize its uncertainty, assumes a data-generating model $p(\boldsymbol{y} \mid \theta)$, and focuses on the posterior distribution $p(\theta \mid \boldsymbol{y})$ for inference
- Posterior distribution is derived from Bayes' theorem
    - $p(\theta \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid \theta) p(\theta)}{p(\mathbf{y})} \propto L(\theta; \mathbf{y}) p(\theta)$.
- We use p for both PDF and PMF (depends on context)
- The $\propto$ trick: $p(\mathbf{y})$ and normalizing constant in $L(\theta; \mathbf{y})$ are dropped, since only θ is random in the posterior.
- For now, assume $p(\theta \mid \mathbf{y})$ is valid PDF and proper, i.e. $\int p(\theta \mid \mathbf{y}) d\theta = 1$

Likelihood

- Likelihood function $L(\theta; \boldsymbol{y})$ characterizes the probability/likelihood
- Function of θ; given $\boldsymbol{y}$ is observed/fixed.

- In MOST cases, the density (or probability mass for discrete data) at observed $\boldsymbol{y}$ is given as
  - $L(\theta; \boldsymbol{y}) = f_{\boldsymbol{Y}}(\boldsymbol{y} \mid \theta)$
- For i.i.d. data observations, $L(\theta; \boldsymbol{y}) = \prod_{i=1}^{n} f_{Y_i}(y_i \mid \theta)$.
- Up to a constant, i.e., $L(\theta; \boldsymbol{y}) = \frac{f_{\boldsymbol{Y}}(\boldsymbol{y}|\theta)}{C}$, as long as C is not involved with θ (can relate to $\boldsymbol{y}$).
- θ with higher likelihood values are more plausible for generating the data, i.e., better approximate for the esitmand.
- **Log-likelihood** is given as $\ell(\theta; \boldsymbol{y}) = \log(L(\theta; \boldsymbol{y}))$.
  - For i.i.d. data observations, $\ell(\theta; \boldsymbol{y}) = \sum_{i=1}^{n} \log(f_{Y_i}(y_i \mid \theta))$.

Maximum likelihood estimation {#mle}

- The **maximum likelihood estimate (MLE)** of an estimand θ is the value $\hat{\theta}$ that maximizes the likelihood function $L(\theta; \vec{y})$.
- The corresponding estimator is the maximum likelihood estimator
  - Use context to determine whether $\hat{\theta}$ refers to the estimate or estimator.
- MLE is usually determined with this procedure:
  1. Determine the likelihood function.
  2. Write the log-likelihood $\ell(\theta) = \log(L(\theta))$. Concept check: Why is this valid?
  3. Take the derivative with respect to θ and set it to 0 to find a critical point $\hat{\theta}$.
  4. Take the second derivative and check that the second derivative evaluated at $\hat{\theta}$ is negative. Concept check: Why?

Ex: Calculate the MLE for p for $Y_1, \ldots, Y_n \overset{i.i.d}{\sim} \text{Geom}(p)$. To find $\hat{p}_{MLE}$, first determine the likelihood. We have

$$L(p; \vec{y}) = \prod_{j=1}^{n} (1-p)^{y_i} p = (1-p)^{\sum_n y_i} p^n.$$

Take the log of the above to find

$$\ell(p) = \sum_{j=1}^{n} y_i \log(1-p) + n \log(p).$$

Take the first derivative

$$\frac{\partial}{\partial p} \ell(p) = -\frac{\sum_{j=1}^{n} y_i}{1-p} + \frac{n}{p}$$

and the second derivative

$$\frac{\partial^2}{\partial p^2} \ell(p) = -\frac{\sum_{j=1}^{n} y_i}{(1-p)^2} - \frac{n}{p^2}.$$

Set the first derivative equal to 0 and solve for the critical point p^* to find

$$p^* = \frac{n}{n + \sum_{j=1}^{n} y_i}$$

- What is an intuitive explanation for the MLE?
- MLE is not "the most likely estimand"
  - Why is this a category error? Recall the frequentist framework
- Out of possible estimand values, the observed data has the highest probability of being generated from the MLE
- Some MLEs that would be helpful to practice with:
  - Binomial
  - Multinomial
  - Poisson
  - Exponential.
- The MLE is invariant
  - $g(\hat{\theta}) = \widehat{g(\theta)}$.

Method of moments estimation {#mom}

- **Method of moments** estimator: match sample moments from the data with analytical/theoretical moments from the model.
  - Sometimes moments will give different estimators.

Ex: Assume our data is distributed i.i.d. Geometric. The mean of a geometric distribution is $\frac{1-p}{p}$. The sample mean is $\hat{Y}$. We can set the two equal to find

$$\frac{1-p}{p} = \bar{Y}.$$

Add 1 to both sides to derive

$$\frac{1}{\hat{p}} = \bar{Y} + 1 \to \hat{p} = \frac{1}{\bar{Y}+1} = \frac{n}{n + \sum Y_i}.$$

Practice problems {#practice}

Problem 1 is sourced from Patrick Dickinson, COL 2022. Problems 2 and 3 are past HW questions.

Problem 1 (True/False questions)

1. There exists an estimator $\hat{\theta}$ for which $\mathrm{MSE}(\hat{\theta}) = (\mathrm{Bias}(\hat{\theta}))^2$
2. If $\hat{\theta}$ is unbiased, then all other estimators are biased.

3. The sample mean is unbiased under all models where a mean exists.
4. The squared error loss of an estimator for its estimand is a random variable.
5. When data comes from a discrete distribution, the likelihood function is also discrete.

Problem 2 (Medical testing)

This problem will be a review of conditioning

We are trying to estimate what proportion of the Harvard body (staff, grad students, and undergraduates) have some unspecified pandemic illness. The administration conducts frequent tests to determine the prevalence and incidence of infection.

(a) Perfect test

Assume that the test is perfect and that whether a person tests positive is distributed i.i.d. Bernoulli with proportion p.

Determine the MLE and produce an estimate of p given the observations. Assume the data is in the vector $\vec{x}$ of length n.

*Note: Why is i.i.d. Bernoulli a poor choice of model here?**

(b) Imperfect test

The **sensitivity** of a test is the probability that someone who has the condition tests positive.

The specificity of a test is the probability that someone who does not have the condition tests negative.

Let sensitivity be a and specificity by b. Correct the likelihood function to accommodate for error in the test.

Problem 3 (DNA sequence)

A DNA sequence can be composed of four different possible base pairs. For example, consider the following sequence:

CTACCTTCAATTGCTGGAACG

(a) Multinomial model

For simplicity, assume that DNA base pairs are selected from a Multinomial distribution (ignore other properties of DNA base pairs).

Let the probabilities corresponding to the base pair selection be represented as $\theta = (p_a, p_c, p_g, p_t)$.

Write the log-likelihood of θ

(b) Markov chain

We can use a Markov chain model to accommodate for possible violations of independence in the genome. Let us model the sequence as a lag-1 Markov model.

In a lag-1 Markov model, we have $\Pr(X_s \mid X_{s-1}, \ldots, X_1) = \Pr(X_s \mid X_{s-1})$.

This model can be represented with a 4×4 transition matrix:

$$T = \begin{pmatrix} \tau_{aa} & \tau_{ac} & \tau_{ag} & \tau_{at} \\ \tau_{ca} & \tau_{cc} & \tau_{cg} & \tau_{ct} \\ \tau_{ga} & \tau_{gc} & \tau_{gg} & \tau_{gt} \\ \tau_{ta} & \tau_{tc} & \tau_{tg} & \tau_{tt} \end{pmatrix}$$

How many parameters does this matrix have?

Assume that the marginal distribution $\Pr(X_1 = x), x \in \{t, c, g, a\}$ is known. Write the log-likelihood of this model.

Problem 3 (MLE and MoM)

Emily and her mom buy a chihuahua from the pet store. The storekeeper, a statistician, mentions that Chihuahua weights can be reasonably modeled as i.i.d. Normal with a standard deviation of one pound. Unfortunately, the storekeeper forgot the average weight of a chihuahua.

(a) MLE

Based on a single observation, y, of her new dog's weight, Emily estimates the mean of the data-generating distribution via maximum likelihood. What estimator does she use? How would Emily estimate $\theta = \mu^2$, where μ is the mean?

(b) MoM

Emily's mom prefers the method of moments. Using the first and second moments respectively, she derives estimators for the same two estimands that Emily estimated. What are Mom's estimators?

(c) MSE

For any estimators that disagree, determine whether Emily or Mom gave the estimator with lower MSE? Can you determine this without actually computing the MSE?

-

Problem 4 (Gaussian mixture model)

Suppose $Y_1, \ldots, Y_n$ are drawn i.i.d. from a Gaussian mixture model, with $n > 1$. In particular, the data come from a standard Normal distribution with probability $\frac{1}{2}$ and otherwise come from a Normal distribution with unknown mean and variance, $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ respectively.

(a) Likelihood calculations

What is the likelihood? What is the log-likelihood?

(b) Arbitrary likelihood

Show that the likelihood can be made arbitrarily large. Hint: Set $\mu = Y_1$.

(c) Non-arbitrary likelihood

Why would we not achieve arbitrarily large likelihood using the same technique if the model were $Y_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$?