STAT 111 Section 01: Review and Introduction

Al Xin

03 Feb 2022

Section Overview

This section will mostly focus on example problems and R. Every week, I will try and anticipate the R help needed for the upcoming homework, i.e., the one due on Friday. The section will not be R office hours, though.

Probability Review

Important Concepts

This is a non-exhaustive list of STAT 110 concepts that will be useful in STAT 111. Some of these are not covered in this week's review, so make sure to check the STAT 110 textbook (available on the STAT 111 Canva) for further review.

- Bayes' rule
- Linearity of expectation
- Fundamental bridge
- Variance of linear combinations
- Sample variance
- Law of large numbers
- Central limit theorem
- Deriving joint, marginal, and conditional distributions

Expectation

If X is a discrete random variable, the **expected value** of X is

$$\mathbb{E}[X] = \sum_x xP(X = x),$$

summing over all the possible values of X.

If X is a continuous random variable with PDF f,

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} xf(x)dx.$$

For both the discrete and continuous case, If the sum diverges, the expectation is undefined.

We use a lot of different notation for expectation; sometimes it's a plain capital letter E, sometimes we use parentheses and not brackets, etc. Generally we have no preference for reasonable variation in notation.

The kth **moment** of X is $\mathbb{E}[X^k]$.

When finding expectation of functions of random variables, we apply the **law of the unconscious statistician (LOTUS)**. Suppose that $Y = g(X)$. The discrete and continuous use of LOTUS (assuming that the expectation of Y exists) are:

$$\mathbb{E}[Y] = \sum_x g(x)P(X = x),$$
$$\mathbb{E}[Y] = \int_{-\infty}^{\infty} g(x)f(x)dx.$$

Linearity of expectation allows us to find the expectation of linear combinations of random variables. For random variables $X_1, \ldots, X_n$ with defined expectations, we have

$$\mathbb{E}\left[a + \sum_{i=1}^{n} b_i X_i\right] = a + \sum_{i=1}^{n} b_i \mathbb{E}[X_i],$$

where $a, b_1, \ldots, b_n$ are constants.

The sample average of realizations of i.i.d. random variables is an unbiased estimator for the expectation.

Variance

If X is a random variable whose first and second moments exist, the variance of X is

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

Taking the variance of linear combinations of random variables, we have

$$\mathrm{Var}\left(a + \sum_{i=1}^{n} b_i X_i\right) = \sum_{i=1}^{n} b_i^2 \mathrm{Var}(X_i) + \sum_{1 \le i < j \le n} 2 b_i b_j \mathrm{Cov}(X_i, X_j).$$

How does the above simplify when variables are independent?

Given $X_1, \ldots, X_n \overset{i.i.d}{\sim}$, we have that the unbiased estimator for the variance is

$$\hat{\sigma}^2 = S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

Law of Large Numbers (LLN)

By the **weak law of large numbers**, when given $X_1, \ldots, X_n \overset{i.i.d}{\sim}$ with finite mean μ and variance σ^2, as $n \to \infty$, we have

$$\bar{X}_n \xrightarrow{p} \mu,$$

where $\bar{X}_n$ is the sample average. Convergence in probability means that $P(|\bar{X}_n - \mu| > \epsilon) \to 0$ as n becomes large.

By the **strong LLN**, in the same setup, we have that

$$\bar{X}_n \overset{a.s.}{\rightarrow} \mu,$$

where convergence almost surely indicates $P(\lim_{n\rightarrow\infty} \bar{X}_n \rightarrow \mu) = 1$.

Central limit theorem

Given $X_1, \ldots, X_n \overset{i.i.d}{\sim}$ with mean μ and variance σ^2, we have

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \overset{d}{\rightarrow} \mathcal{N}(0, 1).$$

Basics of Statistics

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.

Bias, Variance, and Standard Error

The **bias** of an estimator $\hat{\theta}$ for an estimand θ is

$$\mathrm{Bias}_\theta(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta.$$

We say an estimator is unbiased for an estimand if its bias is 0.

Bias describes the *average* difference of an estimator from the estimand, not the error of any particular estimate.Unbiased estimators are not necessarily desirable. We cannot just add or subtract the bias to correct our estimator since the bias depends, in general, on the unknown parameter θ.

The **standard error** of an estimator θ is

$$\mathrm{SE}(\hat{\theta}) = \sqrt{\mathrm{Var}(\hat{\theta})}.$$

Because an estimator is a random variable, it is valid to calculate its variance. Standard error is the same as standard deviation for an estimator, but we use the terminology to distinguish it from the standard deviation of a single observation of the data. Imagine if we could collect the data many times and recalculate the estimate each time. It is this source of variation that standard error measures.

Likelihood

The likelihood function $L(\theta; \boldsymbol{y})$ characterizes the probability/likelihood of observing the data.

- Function of θ; given $\boldsymbol{y}$ is observed/fixed.
- In MOST cases, the density (or probability mass for discrete data) at observed $\boldsymbol{y}$,

$$L(\theta; \boldsymbol{y}) = f_{\boldsymbol{Y}}(\boldsymbol{y}|\theta).$$

- For i.i.d. data observations, $L(\theta; \boldsymbol{y}) = \prod_{i=1}^{n} f_{Y_i}(y_i|\theta)$. - 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, $\ell(\theta; \boldsymbol{y}) = \sum_{i=1}^{n} \log(f_{Y_i}(y_i|\theta))$.

Method of Moments

- MoM: Match the sample moments from data with analytical/theoretical moments from the model; then solve a system of equations to get the estimators.
- First moment: $a_1(\theta) = E_\theta(Y) \longleftrightarrow \bar{\boldsymbol{Y}} = \sum_{i=1}^{n} Y_i = \hat{a}_1$.
- Second moment: $a_2(\theta) = E_\theta(Y^2) \longleftrightarrow \overline{\boldsymbol{Y}^2} = \sum_{i=1}^{n} Y_i^2 = \hat{a}_2$.

Introduction: Causal Inference

Unlike earlier questions of association or prediction, in a causal setting we ask: **How will a change in X's affect the corresponding Y's?**

In the setting we have described in class, our data are in the form of i.i.d. (X_i, Y_i) pairs for $i = 1, \dots, n$. We think of

$$X_i \in \{0, 1\}$$

as the treatment variable, where 1 is if the subject receives some treatment and 0 otherwise (suppose we either change the color of a button on a webpage or we do not). Then we have

$$Y_i = Y_i(X_i) = \begin{cases} Y_i(1) \text{ if } X_i = 1 \\ Y_i(0) \text{ if } X_i = 0 \end{cases} = Y_i(1)X_i + Y_i(0)(1 - X_i).$$

We call $Y(1)$ and $Y(0)$ **potential outcomes**, which are underlying values for some response Y (e.g. whether a person clicks on a button or not) depending on whether they were given the treatment or not. Note that the outcomes are potential because we do not observe them until running the experiment, and even after the experiment we can only ever observe one reality. It is helpful to draw a table of potential outcomes to conceptualize the experimental set-up. As practice, draw two tables, one where there is causation and one without.

The average **causal effect** is

$$\theta = \mathbb{E}Y(1) - \mathbb{E}Y(0) = \mathbb{E}Y(1) - Y(0),$$

while the association is

$$\alpha = \mathbb{E}Y(1)|X = 1 - \mathbb{E}Y(0)|X = 0.$$

In general, the causal effect does not equal the association.

Because X may not be independent of $\{Y(1), Y(0)\}$, we get $\alpha \neq \theta$ in general. If we **randomize** X such that whether subjects receive the treatment is independent of their potential outcomes, then $\alpha = \theta$ meaning we can estimate causation by estimating association.

Introduction: Bayesian Thinking

Suppose θ is our estimand (parameter of interest), Bayesian adopts a prior distribution on θ to characterize its uncertainty, assumes a data-generating model $p(\boldsymbol{y}|\theta)$, and focuses on the posterior distribution $p(\theta|\boldsymbol{y})$ for inference.

Posterior distribution

Posterior distribution contains all the information we need to draw inference from Bayesian perspective.

The posterior can always to be derived by **Bayes' theorem**:

$$p(\theta|\mathbf{y}) = \frac{p(\mathbf{y}|\theta)p(\theta)}{p(\mathbf{y})} \propto L(\theta;\mathbf{y})p(\theta).$$

- We abuse the notation p for both PDF and PMF depending on the 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 this section, we assume $p(\theta|\mathbf{y})$ is valid PDF and proper, i.e. $\int p(\theta|\mathbf{y})d\theta = 1$

Problems

Problem 1 (Cauchy distribution)

(a) Representation of the Cauchy distribution *How is the Cauchy distribution represented?*

(b) Practice with R *Write a function that generates an arbitrary number of observations from a Cauchy random variable using the representation found in (a). Then, plot a histogram of 50, 500, and 1000 observations from your function.*

For future reference, the existing function is `rcauchy`. Try to name a function that doesn't conflict with this one.

(c) Visualize expectation of the Cauchy *In class, we discussed unbiased and consistent estimators. For example, by linearity of expectation, the method of moments estimator is unbiased and consistent. Test the behavior of the MoM estimator for the Cauchy distribution.*

Generate samples of size $100, 200, \ldots, 10000$*. Plot the sample mean of the samples against the sample size in a line plot.*

(d) Method of moments comparison *Using the same sampling scheme, compare the behavior of the method of moments estimator for an Exponential distribution with rate parameter `0.1`. (Review: How does this compare to a standard Exponential distribution? The notation in R can be unintuitive.)*

(e) Challenge: Multiple iterations *Repeat (c), but run multiple simulations of the procedure and plot multiple lines on the graph.*

Problem 2 (Medical diagnosis)

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 COVID-19. The administration conducts frequent tests to determine the prevalence and incidence of infection.

From Jan 26-28, there have been 8,750 tests. There has been a total of 27 positive cases.

(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 accomodate for error in the test.

Problem 3 (Sunspots)

You can see what datasets R has available using the function `data()`

(a) Histogram and ECDF *Plot a histogram and the ECDF of the data* `sunspots` *from R.*

(b) Modeling on monthly sunspot distribution *What distribution does this remind you of? Pick one or more reasonable distributions and compare the ECDFs (number of observations in the sample should equal the number of observations of sunspots).*

(c) MLE *Assume that the data is distributed Exponential with an unknown rate parameter λ. Write the likelihood function and find the MLE.*

(d) MLE of the data *Calculate the estimate of the rate parameter based on the data. Compare the graphs of the histogram and ECDF of an Exponential model and the observations.*

Problem 4 (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 properties of DNA base pair-matching).*

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 $Pr(X_s|X_{s-1}, \ldots, X_1) = Pr(X_s|X_{s-1})$.

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.*