STATS 225: Bayesian Analysis
Supplementary Materials: A brief review of probability [1]

Babak Shahbaba

Department of Statistics, UCI

Winter, 2015

[1] This review is based on Jeffrey Rosenthal's book on rigorous probability

Probability

- We are familiar with statements such as $X \sim \textit{Poisson}(5)$ distribution. We interpret it as X being a non-negative random variable such that $P(X = k) = 5^k \exp(-5)/k!$.
- We call X a *discrete* random variable.
- We are also familiar with statements like $Y \sim \mathrm{Normal}(0, 1)$. It means, for example, $P(a \leq Y \leq b) = \int_a^b \frac{1}{\sqrt{2\pi}} \exp(-\frac{y^2}{2})dy$. We know that $P(Y = y) = 0$ for any real number y. Y is an example of *absolutely continuous* random variable.
- Introductory courses on probability group random variables as either discrete or continuous.
- This is not completely correct. There are other types of random variables. For example, consider a random variable Z defined as follows. We toss a coin, if it's head, we set $Z = X$, otherwise, we set $Z = Y$.

Probability measure

- A mathematical rigorous probability theory is studied in the context of measure theory.

- A probability measure space (or a probability triple), $(\Omega, \mathcal{F}, P)$, defined as follows:
  - Ω is a non-empty set referred to as the sample space (i.e., for example the sample space for the poisson distribution consists of all the non-negative integers).

  - $\mathcal{F}$ is a σ-algebra, which is a collection of measurable (i.e., the probability is defined) subsets of Ω (including Ω itself and the empty set $\emptyset$), all their complements, and their countable unions. That is, Ω is closed under complement (i.e., if $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$), under countable unions and intersections.

  - P is a probability measure mapping between $\mathcal{F}$ and a real number between 0 and 1 such that $P(\emptyset) = 0$, $P(\Omega) = 1$, and P is countably additive, i.e., if $A_1, A2, ...$ are disjoint subsets included in $\mathcal{F}$, we have

$$P(A_1 \cup A_2 \cup, ...) = P(A_1) + P(A_2) +, ...$$

Some additional results

- $P(A^c) = 1 - P(A)$
- Monotonicity: if $A \subseteq B$, where $A, B \in \mathcal{F}$, then $P(A) \leq P(B)$.
- Countable sub-additivity: if $A_1, A_2, ... \in \mathcal{F}$, which may not be disjoint in general, then $P(\bigcup_n A_n) \leq \sum_n P(A_n)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$, where $A, B \in \mathcal{F}$ may not be disjoint.

Some examples

- Tossing a fair coin:
  - $\Omega = \{H, T\}$
  - $\mathcal{F} = \{\emptyset, \Omega, \{H\}, \{T\}\}$
  - $P(\emptyset) = 0, P(\Omega) = 1, P(H) = P(T) = 1/2.$
- Tossing n fair coins:
  - $\Omega = \{(x_1, x_2, ..., x_n)\}$, where $x_i = 01$.
  - $\mathcal{F} = 2^{\Omega} = \{\text{all subsets}\}$
  - $P(A) = \frac{|A|}{2^n}$
- Poisson(5) distribution.
  - $\Omega = \{0, 1, 2, ...\}$
  - $\mathcal{F} = 2^{\Omega} = \{\text{all subsets}\}$
  - $P(A) = \sum_{k \in A} 5^k \exp(-5)/k! \qquad A \in \mathcal{F}$
- What about continuous distributions (where the sample space is not countable) such as Uniform(0, 1)?

Some examples

- The probability triple corresponding to the Uniform(0, 1) distributionin is called the *Lebesgue* measure on [0, 1].
- The sample space is obviously $\Omega = [0, 1]$.
- To construct the corresponding σ-algebra, consider $\mathcal{J}$ as the set of all intervals (e.g., open, closed, half-open, singleton, etc.) contained in [0, 1].
- Now add all the countable unions of intervals, their complements, their countable intersections, etc. (for original intervals and those created later) to create a σ-algebra.
- The smallest σ-algebra create, $\mathcal{B} = \sigma(\mathcal{J})$ is called the *Borel* σ-algebra, and each of its elements is called a *Borel* set.
- For the Uniform(0, 1) distribution, the probability of each interval is equal to the length of that interval. That is, $P([a, b] = P([a, b)) = P((a, b]) = P((a, b)) = b - a$ for $0 \leq a \leq b \leq 1$ (to define P more precisely, we need to discuss *outer measure* and *extension theorem*).
- Similar procedure is used for other continuous distributions.

Random variables

- Random variables: it assigns numerical values to each possible outcome within a sample space, Ω.
- Therefore, given a probability triple, $(\Omega, \mathcal{F}, P)$, a random variable X is a measurable *function* from Ω to the real numbers $\mathcal{R}$.
- For example, we can define $X(Taile) = 0$ and $X(Head) = 1$.
- Since X is measurable, we can talk about $P(X = 0)$ by which we mean $P(Tail)$. In general, we can talk about $P(X \in B)$, for any Borel set B.
- Another example: if $(\Omega, \mathcal{F}, P)$ is a Lebesgue measure on [0, 1], we can define a random variable $X(\omega) = 3\omega + 4$, for all $\omega \in \Omega$.

Random variables

$$
\begin{aligned}
P(X > x) &= P(\omega \in \Omega, X(\omega) > x) \\
&= P\{\omega \in \Omega, 3\omega + 4 > x\} \\
&= P\{\omega > \frac{x-4}{3}\}
\end{aligned}
$$

$$
P(X > x) = \begin{cases} 1 & x \leq 4 \\ \frac{7-x}{3} & 4 \leq x \leq 7 \\ 0 & x \geq 7 \end{cases}
$$

Independence

- Independence: Two events (or random variables) are independent if they do not affect each other's probability. That is, knowing whether an event A has occurred does not change the probability of event B; we say:

$$P(A \cap B)/P(A) = P(B)$$

  or alternatively:

$$P(A \cap B) = P(A)P(B)$$

- We can extend this to any number of events presented as a collection $\{A_\alpha\}_{\alpha \in I}$:

$$P(A_{\alpha_1} \cap A_{\alpha_2} \cap ... \cap A_{\alpha_j}) = P(A_{\alpha_1})P(A_{\alpha_2})...P(A_{\alpha_j})$$

  for any choice of $\alpha_1, \alpha_2, ..., \alpha_j \in I$

Independence

- SImilarly, we can talk about the independence of random variables. Random variables X and Y are independent if

$$P(X \in S_1, Y \in S_2) = P(X \in S_1)P(Y \in S_2)$$

  or alternatively

$$P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y) \qquad \forall x, y \in \mathcal{R}$$

- For a collection of independent random variables, $\{X_\alpha\}_{\alpha \in I}$ we have

$$P(X_{\alpha_i} \in S_i, \forall 1 \leq i \leq n) = \prod_{i=1}^{n} P(X_{\alpha_i} \in S_i)$$

- Note that if X and Y are independent, so are $f(X)$ and $g(Y)$.

Expected values

- For simple random variables whose range is finite, we can represent the distinct values as $x_1, x_2, ..., x_n$ and write $X = \sum_i^n x_i \mathbf{1}_{A_i}$, where $\mathbf{1}_A$ is an indicator function such that

$$\mathbf{1}_A(\omega) = \begin{cases} 1 & \omega \in A \\ 0 & \omega \notin A \end{cases}$$

- The expected value (mean or expectation) for such variables is defined as:

$$\mu_X = E(X) = E\left(\sum_i^n x_i \mathbf{1}_{A_i}\right) = \sum_i^n x_i P(A_i)$$

where $A_i = \{\omega \in \Omega; X(\omega) = x_i\}$, and $\{A_i\}$ is a finite partition (or in general any collection) of Ω.

Expected values

- For example, if we toss a coin and define

$$X(\omega) = \begin{cases} 10 & \text{if Head} \\ 20 & \text{if Tail}, \end{cases}$$

then $E(X) = 10 \times 1/2 + 20 \times 1/2 = 15$.

- Another example: Consider the Lebesgue measure, and let's define X as follows:

$$X(\omega) = \begin{cases} 4 & \omega < 0.25 \\ 6 & \omega = 0.25 \\ 8 & \omega > 0.25, \end{cases}$$

then $E(X) = 4 \times 1/4 + 6 \times 0 + 8 \times 3/4 = 7$

Some properties of expectation

- $E(\mathbf{1}_A) = P(A)$
- $E(c) = c$
- $E(aX + bY) = aE(X) + bE(Y)$
- Expectation is order preserving, i.e., if $X(\omega) \leq Y(\omega)$ for all $\omega \in \Omega$, then $E(X) \leq E(Y)$
- $|E(X)| \leq E(|X|)$
- If X and Y are independent, then $E(XY) = E(X)E(Y)$; note that the other direction does not always hold.

Some properties of expectation

- If $f(X)$ is a function of X, $f : \mathcal{R} \to \mathcal{R}$, then f itself is a simple random variable and can be written as $f(X) = \sum_i^n f(x_i)\mathbf{1}_{A_i}$ and $E(f(x)) = \sum_i^n f(x_i)P(A_i)$.
- Especially, if $f(X) = (x - \mu_X)^2$, the expectation of f is the *variance* of X: $Var(X) = E((x - \mu_X)^2)$, which leads to the well known conclusion that $Var(X) = E(X^2) - E(X)^2$ and also $Var(X) \leq E(X^2)$.

Some other properties of variance

- $Var(aX + b) = a^2 Var(X)$

- $Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$, where $Cov(X, Y) = E((x - \mu_X)(y - \mu_y)) = E(XY) - E(X)E(Y)$

- If X and Y are independent, then $Cov(X, Y) = 0$ and $Var(X + Y) = Var(X) + Var(Y)$.

- Variance of X is in fact its second central moment.

- In general, the k^{th} moment of a random variable is defined as $E(X^k)$.

- With some mathematical precautions, the above properties can be extended to non-simple random variables.

The integration connection

- Similar to the integral, expectation has some nice properties such as linearity, oerder-preserving and so forth.
- In fact, it can be shown that given a probability triple $(\Omega, \mathcal{F}, P)$, and a measurable function X,

$$E(X) = \int_\Omega X dP = \int_\Omega X(\omega) P(d\omega)$$

  which is the integral of X with respect to the probability measure.
- If $(\Omega, \mathcal{F}, P)$ is the Lebesgue measure on [0, 1], and X is Riemann integrable, then the above integral is the common calculus-style integral: $E(X) = \int_0^1 X(t)dt$.
- Even if X is not Riemann integrable (but nevertheless a measurable function with respect to the Lebesgue measure), we can still get the expectation which in this case called the *Lebesgue integral*. That is, the Lebesgue integral is the generalization of the Riemann integral.

Distributions

- Given a random variable X on a probability triple $(\Omega, \mathcal{F}, P)$, its distribution μ is a probability measure on the sample space $\mathcal{R}$ (with the Borel σ-algebra) defined as

$$\begin{aligned} \mu(B) &= P(X \in B) \qquad B \text{ Borel} \\ X &\sim \mu \end{aligned}$$

- Moreover, the *cumulative distribution function* of a random variable X is defined as $F_X(x) = P(X \leq x)$ for $x \in \mathcal{R}$.

- Note that $\lim_{x \to -\infty} F_X(x) = 0$ and $\lim_{x \to \infty} F_X(x) = 1$.

- Recall that we defined the expected value of a measurable function $f(X)$ as $E(f(X)) = \int_\Omega f(X(\omega))P(d\omega)$ with respect to the probability measure P.

- Alternatively, we can define $E(f(X)) = \int_{-\infty}^{\infty} f(t)\mu(dt) = \int_{-\infty}^{\infty} f(t)d\mu$. This is known as the *change of variable theorem*.

Some simple distributions

- One simple distribution is the *point mass* δ_c, which is the distribution of random variable X where $P(X = c) = 1$.
- Another simple distribution is the Poisson(θ) distribution, where $\mu(X) = \sum_{j=0}^{\infty} (\theta^j \exp(-\theta)/j!)\delta_j$
- Normal(0, 1) distribution is defined as $\mu_N(B) = \int_{-\infty}^{\infty} f(t)\mathbf{1}_B(t)\lambda(dt)$. where $f(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$ is called the *density function*, and λ is the Lebesgue measure on R.
- If a distribution has, μ, has a density, f, instead of taking the integral of a function $g(t)$ with respect to μ, we can take the integral $g(t)f(t)$ with respect to λ.

$$\int_{-\infty}^{\infty} g(t)\mu(dt) = \int_{-\infty}^{\infty} g(t)f(t)\lambda(dt)$$

- That is, for such cases we mainly take a calculus-style integral using the density function.

Convergence

- Convergence with probability 1 (almost surely): $P(\lim_{n\to\infty} X_n = X) = 1$.
- Convergence in probability: $\lim_{n\to\infty} P(|X_n - X| \geq \varepsilon) = 0$.
- Convergence in distribution: $\lim_{n\to\infty} P(X_n \leq x) = P(X \leq x)$.
- Convergence with probability 1 $\Rightarrow$ convergence in probability $\Rightarrow$ convergence in distribution.
- Weak law of large numbers: Let $X_1, X_2, ...$ be a sequence of independent random variables with the same mean m and finite variance, then their partial average, $\frac{1}{n}(X_1 + X_2 + ... + X_n)$ converges in probability to m.
- Strong law of large numbers: If besides the above conditions the forth momen, $E((X_i - m)^4)$ is also finite, the partial average converges to m with probability 1 (i.e., almost surely).
- Central limit theorem: Let $X_1, X_2, ...$ be *iid* with finite mean m and finite variance v. Set $S_n = X_1 + X_2 + ... + X_n$. Then as $n \to \infty$, $\frac{S_n - nm}{\sqrt{nv}}$ convergence in distribution to $Z \sim N(0, 1)$.

Conditional probability and expectation

- Conditional probability is simply defined as $P(A|B) = \frac{P(A \cap B)}{P(B)}$, where $P(B) > 0$. This is the proportion of the event B that also includes the event A. In other words, you investigate the event A in a smaller subset of the sample space where the event B occurs.

- Similarly, for random variables, we can define the *conditional distribution* as $P(Y \in S|B) = \frac{P(Y \in S, B)}{P(B)}$.

- Using this new constructed (conditional) distribution, ν, we can define *conditional expectations* in the usual way: $E(Y|B) = \int y \nu d(y)$, $E(f(Y)|B) = \int f(y)\nu(dy)$.

- This seems quite straightforward as long as $P(B) > 0$. But what happens when $P(B) = 0$; for example, can we discuss $P(A|X = 0.5)$ where X is a random variable with Uniform(0, 1) distribution.

Conditional probability and expectation

- To resolve this issue, we regard the conditional probability $P(A|X)$ and expectation $E(Y|X)$ as themselves being random variables that are functions of X. These new values should have the correct expected values:

$$E(P(A|X)) = P(A) \qquad E(E(Y|X)) = E(Y)$$

- However, having the above correct expected values is not enough to specify the distribution of $P(A|X)$ and $E(E(Y|X))$. More specifically, we need these for any Borel $S \subseteq \mathcal{T}$

$$\begin{aligned} E(P(A|X)\mathbf{1}_{X\in S}) &= P(A \cap \{X \in S\}) \\ E(E(Y|X)\mathbf{1}_{X\in S}) &= E(Y\mathbf{1}_{X\in S}) \end{aligned}$$

- Since the above expectations would not be affected by changes on a set of measure 0, the above definitions are only unique up to a set of measure 0. We can in fact change $P(A|X)$ without restriction whenever $P(X = x) = 0$.

Conditional probability and expectation

- Also note that when $S = \mathcal{R}$, we again obtain

$$E(P(A|X)) = P(A) \qquad E(E(Y|X)) = E(Y)$$

- Some useful properties:
  - $P(A \cap B) = P(A)P(B|A) = P(B)P(A|B)$
  - If A is independent of B, then $P(A|B) = P(A)$ and $P(A \cap B) = P(A)P(B)$
  - The total probability rule: if $B_1, B_2, ..., B_n$ partition the sample space (i.e., their are mutually exclusive and $\bigcup_{i=1}^{n} B_i = \Omega$), then

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + ... + P(A|B_n)P(B_n)$$

  - The multiplication rule: If $A_1, A_2, ..., A_n$ is a sequence of events, then $P(A_1 \cap A_2 \cap ... \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2)...P(A_n|A_1 \cap ... \cap A_{n-1})$
- Conditional probabilities play a very important role in Bayesian statistics, and we will discuss them more in future.