A Note on Population or Large Sample Interpretations of Probability Results

Ying Nian Wu, UCLA Statistics

Background materials for STATS courses

Contents

1 As long as you can count and average

Most of the results in probability theory can be understood by middle school students if we interpret them in terms of a population or a large sample. As long as you can count and average, you can understand probability and expectation.

2 Probability and conditional probability

Suppose we randomly sample a person from a population Ω. Let A be the event that the person is male. Then $P(A) = |A|/|\Omega|$, where $|\Omega|$ is the total number of people in the population, A is the sub-population of male people. $P(A)$ is the proportion of male sub-population.

Let B be the event that the person is taller than 6 ft. Then $P(B|A)$ is the proportion of tall people within the male population, thus it should be $|A \cap B|/|B|$, which is $P(A \cap B)/P(B)$. This leads to the definition of conditional probability. It also means that conditional probability is just regular probability where we only consider a sub-population.

We can also think of probability and conditional probability in terms of a large number of independent repetitions. $P(A)$ tells us how often A happens. $P(B|A)$ tells us how often B happens when A happens. That is, probability and conditional probability are frequency and relative frequency respectively.

3 Density function, transformation, expectation

Consider a density function $f(x)$ of a random variable. If we draw a large sample $\{X_i, i = 1, ..., n\}$ from $f(x)$ independently, and if we plot $\{X_i\}$ on the real line. Then we have a distribution of points. The density of the points around a position x is

$$f(x) = \frac{\text{number of points in}(x, x + \Delta x)/n}{\Delta x}.$$

The average of these numbers approaches $\mathrm{E}(X)$. To calculate the average, we can divide the range of X into a large number of small bins, $(x, x + \Delta x)$. The X_i's that fall into this bin are all equal to x approximately. The number of X_i's in this bin is $nf(x)\Delta x$. So the expectation is

$$\mathrm{E}(X) = \frac{1}{n} \sum_x xnf(x)\Delta x = \sum_x xf(x)\Delta x \to \int xf(x)dx,$$

as $\Delta x \to 0$.

Under the transformation $y = h(x)$, we transform the points $\{X_i, i = 1, ..., n\}$ on the x-axis to the points $\{Y_i = h(X_i), i = 1, ..., n\}$ on the y-axis. The density of $\{Y_i\}$ will change according to the slope of $h(x)$. Specifically, suppose h is invertible, so that $x = g(y)$, and h maps $(x, x + \Delta x)$ to $(y, y + \Delta y)$, then

$$\begin{aligned} f_Y(y) &= \frac{\text{number of } Y_i\text{'s in}(y, y + \Delta y)/n}{\Delta y} \\ &= \frac{\text{number of } X_i\text{'s in}(x, x + \Delta x))/n}{\Delta y} \\ &= f_X(x)\Delta x/\Delta y = f_X(g(y))|g'(y)|. \end{aligned}$$

4 Joint, marginal, and conditional

Consider a joint density function $f(x, y)$ for a pair of random variables (X, Y). If we draw a large sample $\{(X_i, Y_i), i = 1, ..., n\}$ from $f(x, y)$, and if we draw a scatterplot by treating each (X_i, Y_i) as a point. Then we have a distribution of points. The density of the points around (x, y) is

$$f(x, y) = \frac{\text{number of points in}(x, x + \Delta x) \times (y, y + \Delta y)/n}{\Delta x \Delta y}.$$

The average of $h(X_i, Y_i)$ is

$$\mathrm{E}[h(X, Y)] = \frac{1}{n} \sum_{x,y} h(x, y)nf(x, y)\Delta x\Delta y = \sum_{x,y} h(x, y)f(x, y)\Delta x\Delta y \to \int h(x, y)f(x, y)dxdy.$$

For the marginal density of X, we can project all the points vertically onto the x-axis, then $f_X(x)$ describes the density of these projected points.

For the marginal density of $X + Y$, we can project all the points onto the x-axis (or the y-axis) but along the direction of $x + y = c$.

For the conditional distribution of Y given $X = x$, we only need to consider the vertical slice $(x, x + \Delta x)$, and look at the distribution of the points within this slice. $\mathrm{E}(Y|X = x)$ is the average of Y_i for all the (X_i, Y_i) within this slice.

Let $h(x) = \mathrm{E}(Y|X = x)$, then $h(x)$ is the regression curve. For instance, if X is age and Y is height, then $h(30)$ is the average heights of 30 years old.

$\mathrm{E}(Y) = \mathrm{E}[h(X)] = \mathrm{E}[\mathrm{E}(Y|X)]$ means that if we change the height of each person to his or her group average, then we are not going to change the sum within the group, thus we are not going to change the overall average. That is, the change $Y \rightarrow h(X)$ won't change the expectation, so that $\mathrm{E}(Y) = \mathrm{E}[h(X)]$. Of course this change will reduce the variance, i.e., $\mathrm{Var}[h(X)] \leq \mathrm{Var}(Y)$, and the difference is the within group variance $\mathrm{E}[\mathrm{Var}(Y|X)]$. Thus $\mathrm{Var}(Y) = \mathrm{E}[\mathrm{Var}(Y|X)] + \mathrm{Var}[\mathrm{E}(Y|X)]$.

5 Acceptance-rejection sampling

Suppose we want to sample from density $f(x)$, which may be a complicated density function. We may sample from a simpler density $g(x)$. Assume that there exists a $C > 1$, such that $Cg(x) \geq f(x)$ for all x, i.e., $Cg(x)$ envelops $f(x)$. Then we can sample X from $g(x)$, and accept X with probability $f(X)/(Cg(X))$. If accept, we return X. Otherwise we go back to sample from $g(x)$ again until the sampled X is accepted.

We can understand the above acceptance-rejection sampling by imagining repeating the above procedure independently. Suppose we sample $\{X_i, i = 1, ..., n\}$ from $g(x)$. Then the number of points in $(x, x + \Delta x)$ is $ng(x)\Delta x$. Among them, only a fraction of $f(x)/(Cg(x))$ is accepted, which amounts to $ng(x)\Delta x f(x)/(Cg(x)) = nf(x)\Delta x/C$ points. The total number of accepted points across all the bins $(x, x + \Delta x)$ is $\sum_x nf(x)\Delta x/C = n/C$. That is, the acceptance rate is $1/C$. Among all the $m = n/C$ accepted points, the proportion of those in the bin $(x, x + \Delta x)$ is $mf(x)\Delta x$. Thus the density of the accepted points is $f(x)$.

Another way to think about it is as follows. Let Ω be the area under the curve $Cg(x)$. Let A be the area under $f(x)$. If we randomly throw n points into Ω. Then the x coordinates of these n points follow $g(x)$. Suppose we accept a point only if it falls into A. Then among all the points that fall into A, the x coordinates of these points is $f(x)$. The acceptance rate is $|A|/|\Omega| = 1/C$, i.e., the ratio of the areas.

6 Markov chain, jump and diffusion

A Markov chain is characterized by (Ω, K, π), and $p^{(t)}$, where Ω is the state space, K is the transition matrix, π is the stationary distribution, and $p^{(t)}$ is the marginal distribution. Specifically, $K(x, y) = P(X_{t+1} = y|X_t = x)$, $p^{(t)}(x) = P(X_t = x)$, and $\pi(x) = P(X_t = x)$ as $t \rightarrow \infty$.

We can imagine 1 million people migrating around the states. $p^{(t)}(x)$ tells us how many people (in the millions) are in state x at time t. $K(x, y)$ tells us the fraction of people in x who will move to y after one step. As people keep moving around, the distribution of the population $p^{(t)}$ will keep changing, and will eventually converge to the stationary distribution π.

The most important formula in Markov chain is $p^{(t+1)}(y) = \sum_x p^{(t)}(x)K(x, y)$. $p^{(t)}(x)$ counts the number (in the millions) of people in x at time t. $K(x, y)$ tells us the fraction of those in x who will move to y, so the number of people from x to y is $p^{(t)}(x)K(x, y)$. Summing over all the possible sources x, we get the number of people in y at time $t + 1$. At stationarity we have $\pi(y) = \sum_x \pi(x)K(x, y)$. In matrix notation, if we treat $p^{(t)}$, π as row vectors, then $p^{(t+1)} = p^{(t)}K$, and $\pi = \pi K$.

Such a population migration interpretation is perfect for Google's page rank, which computes $\pi = p^{(0)}K^t$ for large t. $\pi(x)$ counts how many people are in x, i.e., how popular the webpage x is.

In real life, people move in continuous time. Just like making a movie, we can discretize the time into small intervals such as $(t, t + \Delta t)$, which is the time period between two consecutive frames. We can model what happens within each $(t, t + \Delta t)$, i.e., the transition matrix $K^{(\Delta t)}$. For $i \neq j$, we can model $K^{(\Delta t)}_{ij} = a_{ij}\Delta t$, so $K^{(\Delta t)}_{ii} = 1 - \sum_{j \neq i} a_{ij}\Delta t = 1 + a_{ii}\Delta t$, thus $K^{(\Delta t)} = I + A\Delta t$.

We can still use $p^{(t)}$ to denote the marginal distribution of the population. Then the population distribution $p^{(t)}$ changes continuously if we show the movie, and $p^{(t)} \to \pi$. In fact, $dp^{(t)}/dt = p^{(t)}A$.

In the above, we assume the state space Ω is finite and discrete, so that people jump from one state (webpage) to another. In diffusion, the state space is continuous, such as a real line. Let X_t be the position at time t, we may model the movement of a particle within a small period between two consecutive frames as $X_{t+\Delta t} = X_t + \sigma\epsilon_t\sqrt{\Delta t}$, where ϵ_t are iid random variables and $\text{E}(\epsilon_t) = 0$, $\text{Var}(\epsilon_t) = 1$.

Imagine that you pour a drop of milk into your cup of coffee, the drop of milk will diffuse to the whole cup. The drop of milk is a population of particles, and X_t is the position of a random particle. The distribution of the drop of milk is $p^{(t)}$. The scaling $\sqrt{\Delta t}$ is crucial, because it ensures that $\text{Var}(X_t|X_0 = x) = \sigma^2 t$, which is independent of Δt. That is, if you make a movie of the drop of milk, you will see the same movie no matter whether the Δt of your video camera is 1/24 second or 1/100 second.

In the case of diffusion, A is a verb, which is $\frac{\sigma^2}{2}\frac{\partial^2}{\partial x^2}$. The continuous change of the drop of milk follows the partial differential equation

$$\frac{\partial p^{(t)}(x)}{dt} = \frac{\sigma^2}{2}\frac{\partial^2 p^{(t)}(x)}{\partial x^2}.$$

7 Markov chain Monte Carlo

In MCMC, we want to sample from a target distribution $\pi(x)$, by an iterative algorithm that can be modeled by a Markov chain.

In the Metropolis algorithm, you can use $K(x, y)$ to drive a base chain. $K(x, y)$ is like a proposal. But if the population migration is driven by K, it may not reach the stationary distribution π. Even if the current distribution is π, the distribution in the next step may not be π. The minimal requirement is that at least we should maintain π if we are already in π. Then we can hope that the population will converge to π even if we start from an arbitrary distribution.

Now let us consider how to maintain the stationary distribution. Suppose the population is in stationarity. Then the number of people in x is $\pi(x)$. According to K, the number of people in x who plan to go to y is $\pi(x)K(x, y)$ (e.g., 40 people). The number of people in y who plan to go to x is $\pi(y)K(y, x)$ (e.g., 50 people). This may create imbalance in the trade of people. To maintain the balance, we may only allow $\min(\pi(x)K(x, y), \pi(y)K(y, x))$ (e.g., 40 people) to go through in each direction. That is, we grant the visa applications of those who are in x and who plan to go to y with probability

$$\frac{\min(\pi(x)K(x, y), \pi(y)K(y, x))}{\pi(x)K(x, y)} = \min\left[1, \frac{\pi(y)K(y, x)}{\pi(x)K(x, y)}\right].$$

This ensures the detailed balance between every pair of states x and y, and maintains the stationary distribution.

For the Gibbs sampler, we sample from a joint distribution $\pi(x, y)$ by sampling X conditional on the current value of Y, and then sampling Y conditional on the current value of X. In general, if we want to sample from a joint distribution $\pi(x_1, ..., x_k, ..., x_d)$, then for k in $1, ..., d$, we sample $\pi(x_k|x_{-k})$ where x_{-k} is the current values of all the components excluding x_k.

For instance, let Ω be a two dimensional island, and π be the uniform distribution over Ω. The Gibbs sampler is such that you first randomly relocate horizontally, and then randomly relocate vertically. Consider 1 million people starting from the same spot on the island. If everyone moves independently according to the Gibbs sampler, then the distribution of the people will eventually become uniform over the whole island, and the distribution will not change even if we keep shuffling people's positions.

For a general $\pi(x, y)$, you can imagine 1 million points whose distribution follows $\pi(x, y)$. Then $\pi(x|y)$ is the distribution of points in the horizontal slice $(y, y + \Delta y)$, and $\pi(y|x)$ is the distribution of points in the vertical slice $(x, x + \Delta x)$. If each point moves by following the Gibbs sampler, then within each horizontal slice $(y, y + \Delta y)$, we let the points randomly relocate, but according to the distribution $\pi(x|y)$. After the relocation, the distribution of the points in the horizontal slice $(y, y + \Delta y)$ won't change, so the overall distribution of the points is still $\pi(x, y)$. The same with random relocation in the vertical slices. Therefore, the Gibbs sampler maintains π, and is expected to converge to π even if we start from an arbitrary distribution.

8 Bayes rule and Bayes network

The above considerations of the Markov chains are all forward in thinking, i.e., we want to know what happen as time goes by. We can also think in the inverse direction. Consider three time points 0, s, and T, where $s < T$. For a random person, suppose I tell you $X_0 = a$ and $X_T = b$. I then ask you what is X_s or what is the probability $X_s = c$? This is a typical problem of Bayes reasoning.

You can imagine 1 million people start from state a at time 0. Then you can translate the above question into: Among all the people who end up in state b at time T, how many people were at state c at time s?

Now let us count. The distribution of people at time s is $p^{(s)}$. The distribution of people at time T is $p^{(T)}$. Let $M(x, y) = P(X_T = y|X_s = x)$, which is the fraction of people in x at time s who will end up in y at time T. Then the number of people who were in c at time s and who end up in b at time T is $p^{(s)}(c)M(c, b)$. Thus among all the $p^{(T)}(b)$ people who end up in b at time T, the number of people who were in c at time s is $p^{(s)}(c)M(c, b)/p^{(T)}(b)$. This is an example of Bayes rule. The recursion $p^{(t+1)} = p^{(t)}K$ helps us do the calculations, and can be generalize to belief propagation for Bayes network.

More specifically, we can define $\Pi^{(t)}(x) = P(X_t = x|X_0 = a)$. Then $\Pi^{(t+1)}(y) = \sum_x \Pi^{(t)}(x)K(x, y)$. We can define $\Lambda^{(t)}(x) = P(X_T = b|X_t = x)$. Then $\Lambda^{(t)}(x) = \sum_y \Lambda^{(t+1)}(y)K(x, y)$. Then $P(X_s = c|X_0 = a, X_T = b) = P(X_s = c|X_0 = a)P(X_T = b|X_s = c, X_0 = a)/P(X_T = b|X_0 = a) \propto \Pi^{(s)}(c)\Lambda^{(s)}(c)$. We can normalize the product to get the probability. Π stands for prior. Λ stands for likelihood. $\Pi^{(t)}(x)$ tells us among all the people at state a at time 0, what is the fraction who come to state x at time t. $\Lambda^{(t)}(x)$ tells us among all the people in state x at time t, what is the fraction who end up in state b at time T. The recursive formulas can be understood by counting the population. $\Pi^{(s)}(c)\Lambda^{(s)}(c)$ counts the number of people going through $a \to c \to b$ at times 0, s, and T. $1/\alpha = \sum_x \Pi^{(s)}(x)\Lambda^{(s)}(x)$ counts the number of people going through $a \to b$ at times 0 and T. Then $\alpha\Pi^{(s)}(c)\Lambda^{(s)}(c)$ tells us among all the people who start from a and end up in b, what is the fraction going through c at time s. In terms of computation, we can compute $\Pi^{(t)}$ for $t = 0, 1, ..., s$, and compute $\Lambda^{(t)}$ for $t = T, T - 1, ..., s$. This type of recursive computation underlies the forward-backward algorithm in hidden Markov model. It is also related to dynamic programming or Viterbi algorithm if we replace sum by max.

In Bayesian network, consider a medical diagnosis example, X_0 is like the background information of the patient, X_T is like the medical test results of the patient, and X_s is like the diseases the patient may have.