Math 30530, Introduction to Probability

Spring 2019

Notes on discrete random variables

Random variables: Given an experiment that leads to a sample space S of possible outcomes, and a probability function P that assigns probabilities to events, a *random variable* X is an assignment of a number to each possible outcome (think of it as a filter that assigns to each outcome a single number that extracts from that outcome a particular piece of relevant information). We write

$$X : S \to \mathbb{R}$$

to indicate that X is a function from the sample space to the reals.

Probability mass function: If X is *discrete* — its range, its set of possible values, can be listed as $\{x_1, x_2, \ldots\}$ — then we can completely encode X by its *probability mass function*. This is the function

$$p_X : \text{Range}(X) \to [0, 1]$$

that's given by $p_X(x_i) = P(X = x_i)$; so the probability mass function is really nothing more than a list of the possible values that X can take on, together with information about the probability with which it takes on that value. We say that two random variables X and Y are *equivalent* (or sometimes just *the same*) if they have the same probability mass functions — so, same range of possible values, and same probabilities of taking on those values. Two very different experiments may lead to the same (equivalent) random variables.

Convention: For the rest of this note, all random variables are discrete.

Expectation: We define the *expectation* of X (also called the *mean*) to be

$$E(X) = \sum_{x_i} x_i P(X = x_i) = \sum_{x_i} x_i P_X(x_i).$$

$E(X)$ is a measure of the average or typical value of X, over many repetitions of the experiment. It's computed as a weighted average of all possible values that X can take on, with each value weighted by the probability that X takes on that value. $E(X)$ need not be an actual value that X can take on. Notice that equivalent random variables have the same expectation, so it makes sense to talk about the *expectation of a mass function*.

If S, the sample space, is discrete ($S = \{s_1, s_2, s_3, \ldots\}$) then

$$E(X) = \sum_{s} X(s) P(\{s\}).$$

To see that $\sum_s X(s)P(\{s\})$ is the same as $\sum_{x_i} x_i P(X = x_i)$, just group together the s's according to what their value is under X.

If X takes on (some subset of) the values $\{0, 1, 2, \ldots\}$, then

$$E(X) = P(X \geq 1) + P(X \geq 2) + P(X \geq 2) + \cdots .$$

$E(X)$ is sometimes written as μ or μ_1.

Variance: We define the *variance* of X to be

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

$\mathrm{Var}(X)$ is a measure of on average how far values of X are from the expectation/mean. Notice that $(X - E(X))^2$ is a perfectly reasonable random variable — assigning value $(X(s) - \mu)^2$ to s — so the definition makes sense. Notice also that equivalent random variables have the same variance.

The variance can be calculated by any of the following formulae:

$$\begin{aligned} \mathrm{Var}(X) &= \sum_{x_i} (x_i - \mu)^2 P_X(x_i) \\ &= \sum_s (X(s) - \mu)^2 P(\{s\}) \\ &= E(X^2) - E(X)^2 \end{aligned}$$

where $E(X^2) = \sum_{x_i} x_i^2 P_X(x_i)$, sometimes written as μ_2.

The variance is always non-negative. If X is measured in *units* then $\mathrm{Var}(X)$ is measured in *units squared*. The *standard deviation* of X, which is measured in *units*, is defined to be $\sqrt{\mathrm{Var}(X)}$.

$\mathrm{Var}(X)$ is sometimes written as σ^2, and the standard deviation as σ.

Probability generating function: If X has range that is a subset of $\{0, 1, 2, 3, \ldots\}$, then we can completly encode the probability mass function of X by its *probability generating function*. This is the function

$$G_X : [0, 1] \to \mathbb{R}$$

that's given by

$$G_X(x) = p_X(0) + p_X(1)x + p_X(2)x^2 + p_X(3)x^3 + p_X(4)x^4 + \cdots .$$

If the probability generating function can be re-expressed in some simpler form than just the long sum above, it is a useful tool for calculating expectation and variance:

$$E(X) = G'_X(1) \ \left(= \left.\frac{dG_X}{dx}\right|_{x=1} \right), \qquad \mathrm{Var}(X) = G''_X(1) + G'_X(1) - (G'_X(1))^2.$$

The seven basic discrete mass functions: Here is a list of the seven commonly occurring discrete mass functions. For each, I've listed its name, abbreviation, typical usage, parameters, range, mass function, expectation, and, where it will be useful to know, its variance and/or probability generating function. I've also added a few special properties of some of the mass functions.

Uniform Uni($\{x_1, x_2, \ldots, x_n\}$) or Uniform($\{x_1, x_2, \ldots, x_n\}$) — Models selection of a number uniformly at random from a finite set $\{x_1, x_2, \ldots, x_n\}$.

- **Parameters**: $x_1, x_2, \ldots, x_n$
- **Range**: $x_1, x_2, \ldots, x_n$
- **Mass function**: $p_X(x_1) = \frac{1}{n}$ for each i
- **Expectation**: $\frac{x_1+x_2+\cdots+x_n}{n}$; in the special case where set is $\{1, 2, \ldots, n\}$, expectation is $\frac{n+1}{2}$
- **Variance**: In the special case where set is $\{1, 2, \ldots, n\}$, variance is $\frac{n^2-1}{12}$

Bernoulli Ber(p) or Bernoulli(p)— Models the success or failure of a single trial, where the probability of success is p. Traditionally the failure probability $1 - p$ is written as q.

- **Parameter**: p
- **Range**: $0, 1$ (0 indicates failure, 1 indicates success)
- **Mass function**: $p_X(0) = q$ and $p_X(1) = p$
- **Expectation**: p
- **Variance**: pq
- **Generating function**: $q + px$

Binomial Bin(n, p) or Binomial(n, p)— Models the number of successes, when a Bernoulli(p) trial is repeated independently n times.

- **Parameters**: n, p
- **Range**: $0, 1, \ldots, n$
- **Mass function**: $p_X(k) = \binom{n}{k} p^k q^{n-k}$ for each k
- **Expectation**: np
- **Variance**: npq
- **Generating function**: $(q + px)^n$

Geometric Geom(p) or Geometric(p) — Models the number of independent Bernoulli(p) trials performed until first success is recorded.

- **Parameter**: p
- **Range**: $1, 2, \ldots, \infty$
- **Mass function**: $p_X(k) = q^{k-1}p$ for $k \neq \infty$; $p_X(\infty) = 0$
- **Expectation**: $\frac{1}{p}$
- **Variance**: $\frac{q}{p^2}$
- **Generating function**: $\frac{px}{1-qx}$
- **Special property**: The geometric is the only discrete random variable that is *memoryless*: for $n, m > 0$,

$$P(X > n + m | X > m) = P(X > n)$$

Negative Binomial NegBin(k, p) or NegativeBinomial(k, p) — Models the number of independent Bernoulli(p) trials performed until kth success is recorded.

- **Parameters**: p, k
- **Range**: $k, k+1, k+2, \ldots, \infty$
- **Mass function**: $p_X(n) = \binom{n-1}{k-1} q^{n-k} p^k$ for $n \neq \infty$; $p_X(\infty) = 0$
- **Expectation**: $\frac{k}{p}$
- **Variance**: $\frac{kq}{p^2}$
- **Generating function**: $\left(\frac{px}{1-qx}\right)^k$ (we didn't derive this; it is not yet relevant)

Hypergeometric HGeom(r, g, n) or Hypergeometric(r, g, n)— Models the number of red balls selected in a random sample of size n from a bag with r red balls and g green balls, $n \leq r, g$.

- **Parameters**: n, r, g
- **Range**: $0, 1, \ldots, n$
- **Mass function**: $P_X(k) = \frac{\binom{r}{k}\binom{g}{n-k}}{\binom{r+g}{n}}$ for each k
- **Expectation**: $\left(\frac{r}{r+g}\right) n$ (same as it would be if experiment was Binomial $\left(n, \frac{r}{r+g}\right)$, i.e., if balls were selected one at a time *with replacement*)
- **Variance**: Not relevant to us at the moment; but for large r, g, it is very close to $\left(\frac{rg}{(r+g)^2}\right) n$ (what it would be if experiment was Binomial $\left(n, \frac{r}{r+g}\right)$)
- **Generating function**: Not relevant

Poisson Poi(λ) or Poisson(λ) — Models the number of occurrences of a rare event in unit time, when there are on average λ occurrences per unit time, when occurrences in disjoint time periods are independent, and when simultaneous occurrences are very unlikely.

- **Parameter**: λ
- **Range**: $0, 1, 2, \ldots$
- **Mass function**: $p_X(k) = \frac{\lambda^k}{k!} e^{-\lambda}$ for each k
- **Expectation**: λ
- **Variance**: λ
- **Generating function**: $e^{-\lambda} e^{\lambda x}$
- **Special property**: Poisson(λ) is a very good approximation for Binomial $\left(n, \frac{\lambda}{n}\right)$ for large n and fixed λ; more generally it is a very good approximation for Binomial(n, p) when n is large, p is small, and $np \approx \lambda$