STAT 516: Basic Probability and its Applications

Lecture 4: Random variables

Prof. Michael Levine

September 15, 2015

What is a random variable?

- Often, it is hard and/or impossible to enumerate the entire sample space
- For a coin flip experiment, the sample space is $\mathcal{S} = \{H, T\}$
- Define a function X s.t. $X(T) = 0$ and $X(H) = 1$
- X maps the sample space onto the space $\mathcal{D} = \{0, 1\}$

Formal definition

- A function X that assigns to each element of $s \in \mathcal{S}$ one and only one number $X(s) = x$ is a **random variable**
- The **space** or **range** of X is the set of real numbers $\mathcal{D} = \{x : x = X(s), s \in \mathcal{S}\}$.
- A random variable is **discrete** if its range $\mathcal{D}$ is countable
- A random variable is **continuous** if its range D is an interval of real numbers

Discrete random variable example

- The quality control process: we sample batteries (or any other industrially manufactured product) as it comes off the conveyor line. Let F denote the faulty and S the good one. The sample space is $\mathcal{S} = \{S, FS, FFS, \ldots\}$. Let X be the number of batteries that is examined before the experiment stops. The, $X(S) = 1, X(FS) = 2, \ldots$.

Probability mass function

- Let X have the range $\mathcal{D} = \{d_1, \ldots, d_m\}$
- The induced probability $p_X(d_i)$ on $\mathcal{D}$ is

$$p_X(d_i) = P[\{s : X(s) = d_i\}]$$

  for $i = 1, \ldots, m$
- $p_X(d_i)$ is the **probability mass function (pmf)** of X
- For any subset $D \in \mathcal{D}$ the induced probability distribution is

$$P_X(D) = \sum_{d_i \in \mathcal{D}} p_X(d_i)$$

- It is easy to verify that $P_X(D)$ is a probability on D

Example: first roll in the game of craps

- Sample space $\mathcal{S} = \{(i,j) : 1 \le i,j \le 6\}$ and $P[\{i,j\}] = \frac{1}{36}$
- The random variable is $X(i,j) = i + j$ with the range $\mathcal{D} = \{2, 3, \ldots, 12\}$
- Easy to put together a pmf of X in the table form
- Check that e.g. for $B_1 = \{x : x = 7, 11\}$ $P_X(B_1) = \frac{2}{9}$

Continuous case

- We assume that for any $(a, b) \in \mathcal{D}$ there exists a function $f_X(x) \geq 0$ s.t.

$$P_X[(a, b)] = P[\{s \in \mathcal{S} : a < X(s) < b\}] = \int_a^b f_X(x)\, dx$$

- We also require that $P_X(\mathcal{D}) = \int_{\mathcal{D}} f_X(x) = 1$
- $f_X(x)$ is a **probability density function** or **pdf**

Example

- Choose a *random* number from $(0,1)$
- Sensible assumption would be

$$P_X[(a,b)] = b - a$$

for $0 < a < b < 1$

- The pdf of X is

$$f_X(x) = \begin{cases} 1 & 0 < x < 1 \\ 0 & \text{elsewhere} \end{cases}$$

- Any probability can now be readily computed

Cumulative distribution function

- For a random variable X a **cumulative distribution function** or **cdf** is

$$F_X = P_X((-\infty; x]) = P(\{s \in \mathcal{S} : X(s) \leq x\})$$

- The short notation is $P(X \leq x)$
- For discrete random variables a cdf is a **step function**

Example: a geometric random variable

- Starting at a fixed time, we observe the gender of each newborn child at a hospital until a boy is born. Let $p = P(B)$ and X the number of births observed until "success"
- Then,
$$p_X(x) = (1-p)^{x-1}p$$
for $x = 1, 2, 3 \ldots$
- Verify that
$$F_X(x) = 1 - (1-p)^x$$
for any positive integer x
- More generally,
$$F_X(x) = \begin{cases} 0 & x \leq 1 \\ 1-(1-p)^{[x]} & x \geq 1 \end{cases}$$
where $[x]$ is the **integer part** of x

Computation

- What is the probability that we have to wait no more than 5 times for the birth of a boy? Assume $p = 0.51$
- Use the following R command: $pgeom(q = 5, prob = 0.51)$; the result is 0.9718
- On the contrary, the probability of having to wait more than 3 times is $1 - pgeom(q = 3, prob = 0.51)$

A median of X

- Since a cdf of a discrete random variable is a step function, it does not attain all possible values of X
- How, in general, do we split the distribution into two halves?
- Any number m such that $P(X \leq m) \geq 0.5$ and also $P(X \geq m) \geq 0.5$ is called a **median** of F (or of X)
- The median need not be unique

Characterization of a median of X

- Let X be a random variable with the CDF $F(x)$. Let m_0 be the first number such that $F(m_0) \geq 0.5$ and m_1 the last number such that $P(X \geq x) \geq 0.5$. Then, m is a median of X if and only if $m \in [m_0, m_1]$
- The proof uses the right continuity of a cdf

Example

- X - a random number on $(0, 1)$
- Check that

$$F_X(x) = \begin{cases} 0 & x < 0 \\ x & 0 \leq x < 1 \\ 1 & x \geq 1 \end{cases}$$

Equality in distribution

- X and Y are **equal in distribution** or $X \overset{D}{=} Y$ iff

$$F_X(x) = F_Y(x)$$

  for all $x \in \mathbb{R}$
- This is non-trivial: compare X from the last example and $Y = 1 - X$

Main Properties of cdfs

- F is non-decreasing
- $\lim_{x \to -\infty} F(x) = 0$
- $\lim_{x \to \infty} F(x) = 1$
- $F(x)$ is right continuous

Some other important properties of cdf

- For $a < b$,
$$P[a < x \leq b] = F_X(b) - F_X(a)$$
- For any random variable
$$P(X = x) = F_X(x) - F_X(x-)$$
for any $x \in \mathbb{R}$

Example

- X is a lifetime in years of a mechanical part
- $$F_X(x) = \begin{cases} 0 & x < 0 \\ 1 - \exp(-x) & x \geq 0 \end{cases}$$
- $$f_X(x) = \begin{cases} \exp(-x) & 0 < x < \infty \\ 0 & \text{elsewhere} \end{cases}$$
- $P(1 < X \leq 3) = F_X(3) - F_X(1) = \exp(-1) - \exp(-3) = 0.318$