3.1–3.3 Binomial Distribution and Discrete Random Variables

Prof. Tesler

Math 186
Winter 2017

Random variables

- A *random variable* X is a function assigning a real number to each outcome in a sample space.
- A biased coin has probability p of heads, $q = 1 - p$ of tails. Flip the coin 3 times and let X denote the number of heads:
$$X(HHH) = 3 \quad X(HHT) = X(HTH) = X(THH) = 2$$
$$X(TTT) = 0 \quad X(HTT) = X(THT) = X(TTH) = 1$$
- The *range of X* is $\{0, 1, 2, 3\}$.
- The discrete *probability density function (pdf)* is $p_X(k) = P(X = k)$:
$$p_X(0) = q^3 \qquad p_X(1) = 3pq^2 \qquad p_X(2) = 3p^2q \qquad p_X(3) = p^3$$
- $p_X(k)$ is defined for *all* real numbers k. In this case, $p_X(k) = 0$ for $k \neq 0, 1, 2, 3$:
$$p_X(4) = 0 \qquad p_X(2.5) = 0 \qquad p_X(-3) = 0 \qquad p_X(\pi) = 0 \qquad \ldots$$

Discrete random variables

- In the preceding example, the range of X is a *discrete set*, not a continuum (such as the real number interval $[0, 3]$).
  So X is a *discrete random variable*.

- Sometimes it's called a *probability mass function* (pmf) in the discrete case, vs. a *probability density function* (pdf) in the continuous case. We'll use *probability density function* for both.

- **Notation $\boldsymbol{p_X(k) = P(X = k)}$:** Use capital letters (X) for random variables and lowercase (k) to stand for numeric values.

- A discrete probability density function requires $p_X(k) \geqslant 0$ for all k, and that the total probability is $\sum_k p_X(k) = 1$. On the previous slide:

$$\sum_k p_X(k) = p_X(0) + p_X(1) + p_X(2) + p_X(3)$$
$$= q^3 + 3pq^2 + 3p^2q + p^3 \qquad = (q+p)^3 = 1^3 = 1$$

Binomial distribution

- A biased coin has probability p of heads, $q = 1 - p$ of tails.
- Flip the coin 7 times.
- $P(HHTHTTH) = ppqpqqp = p^4q^3 = p^{\text{\# heads}}q^{\text{\# tails}}$
- $P(\text{4 heads in 7 flips}) = \binom{7}{4}p^4q^3$
- Flip the coin n times ($n = 0, 1, 2, 3, \ldots$).
  Let X be the number of heads.
  The *probability density function (pdf)* of X is

$$p_X(k) = P(X = k) = \begin{cases} \binom{n}{k}p^kq^{n-k} & \text{if } k = 0, 1, \ldots, n; \\ 0 & \text{otherwise.} \end{cases}$$

- **Interpretation:** Repeat this experiment (flipping a coin n times and counting the heads) a huge number of times. The fraction of experiments with $X = k$ will be approximately $p_X(k)$.

Binomial distribution

$$p_X(k) = P(X = k) = \begin{cases} \binom{n}{k} p^k q^{n-k} & \text{if } k = 0, 1, \ldots, n; \\ 0 & \text{otherwise.} \end{cases}$$

- The range of X is $\{0, 1, 2, \ldots, n\}$.
- $p_X(k) \geqslant 0$ for all values k.
- The sum of all probability densities is 1:

$$\sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} = (p+q)^n = 1^n = 1$$

- The relationship to the binomial formula is why it's named the *binomial distribution*.

Genetics example

- Consider pea plants from a $Tt \times Tt$ cross. The offspring have

Genotype	**Probability**	**Phenotype**
TT	1/4	tall
Tt	1/2	tall
tt	1/4	short

  so the phenotypes have $P(\text{tall}) = 3/4$, $P(\text{short}) = 1/4$.

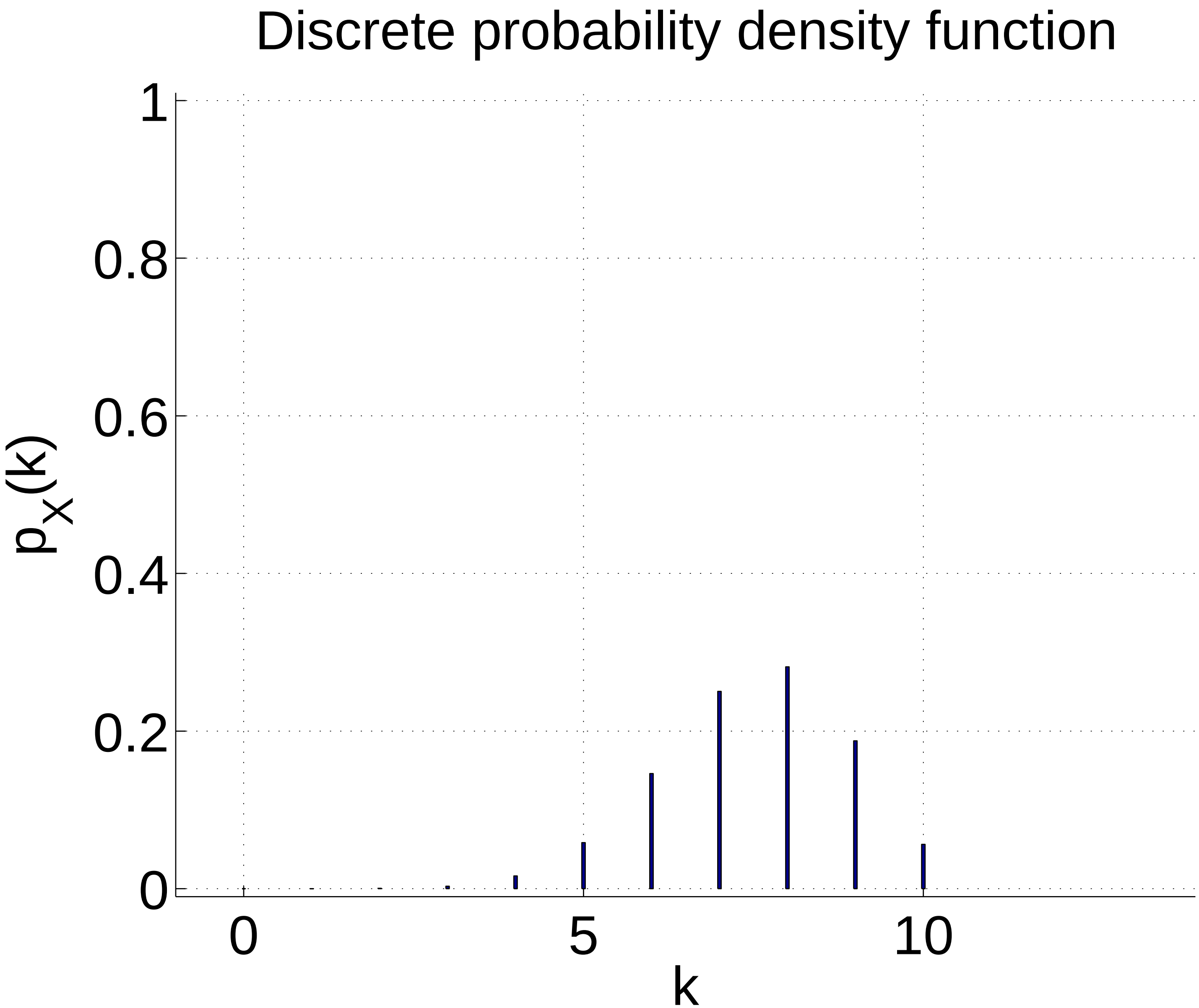
- If there are 10 offspring, the number X of tall offspring has a binomial distribution with $n = 10$, $p = 3/4$:

$$p_X(k) = P(X = k) = \begin{cases} \binom{10}{k} (3/4)^k (1/4)^{10-k} & \text{if } k = 0, 1, \ldots, 10; \\ 0 & \text{otherwise.} \end{cases}$$

- **Later:** We will see other bioinformatics applications that use the binomial distribution, including genome assembly and Haldane's model of recombination.

Binomial distribution for $n = 10$, $p = 3/4$

k	pdf
0	0.00000095
1	0.00002861
2	0.00038624
3	0.00308990
4	0.01622200
5	0.05839920
6	0.14599800
7	0.25028229
8	0.28156757
9	0.18771172
10	0.05631351
other	0



Cumulative Distribution Function (cdf)

- The *Cumulative Distribution Function (cdf)* of random variable X is

$$F_X(k) = P(X \leqslant k)$$

  defined over *all* real numbers k.

- In our example,

$$\begin{aligned} F_X(1) &= P(X \leqslant 1) = p_X(0) + p_X(1) \\ &= 0.00000095 + 0.00002861 = 0.00002956 \end{aligned}$$

$$\begin{aligned} F_X(2) &= P(X \leqslant 2) = p_X(0) + p_X(1) + p_X(2) \\ &= 0.00000095 + 0.00002861 + 0.00038624 = 0.00041580 \end{aligned}$$

Alternately:

$$\begin{aligned} &= F_X(1) + p_X(2) \\ &= .00002956 + 0.00038624 = 0.00041580 \end{aligned}$$

CDF in-between points with nonzero probability

- Note that

$$F_X(1.5) = P(X \leqslant 1.5) = p_X(0) + p_X(1) = F_X(1)$$

- The binomial distribution has nonzero probability only at integers.
- In-between integers,
  - PDF: $p_X(k) = 0$
  - CDF: $F_X(k) = F_X(\lfloor k \rfloor)$,

    where $\lfloor k \rfloor$ is the *floor of k* (largest integer $\leqslant k$):

$$\lfloor 3 \rfloor = 3, \quad \lfloor -3 \rfloor = -3, \quad \lfloor 3.2 \rfloor = 3, \quad \lfloor -3.2 \rfloor = -4.$$

Warning

Be careful, this is just our first example.
If the range of a random variable includes non-integer locations, go down to the largest value $\leqslant k$ with nonzero probability instead of to $\lfloor k \rfloor$.

CDF outside of the range

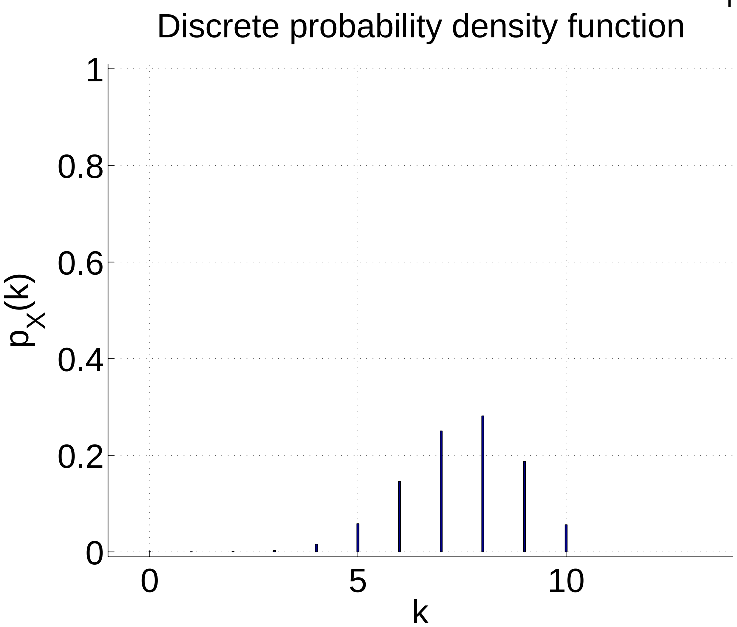
- In this example, the range of X is $\{0, 1, \ldots, 10\}$.
- $F_X(-3.2) = P(X \leqslant -3.2) = 0$ since minimum X in range is 0.
- $F_X(12.8) = P(X \leqslant 12.8) = 1$ since the whole range is $\leqslant 12.8$.
- This example has a bounded range.
  $F_X(k) = 0$ below the range and $F_X(k) = 1$ above the range.
  But not all random variables have a bounded range.
  Instead, for any random variable, we have asymptotic results:

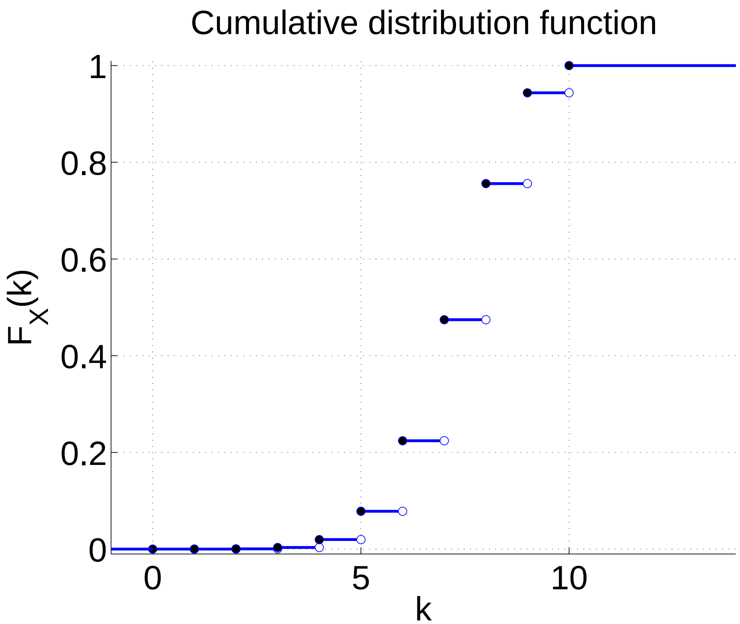
$$\lim_{k \to -\infty} F_X(k) = 0 \qquad \lim_{k \to +\infty} F_X(k) = 1$$

- As k goes from $-\infty$ to ∞, the cdf weakly increases.
- For a discrete random variable, the cdf jumps where the pdf is nonzero.

Binomial distribution for $n = 10$, $p = 3/4$

k	pdf $p_X(k)$		cdf $F_X(k)$
		$k < 0$	0
0	0.00000095	$0 \leqslant k < 1$	0.00000095
1	0.00002861	$1 \leqslant k < 2$	0.00002956
2	0.00038624	$2 \leqslant k < 3$	0.00041580
3	0.00308990	$3 \leqslant k < 4$	0.00350571
4	0.01622200	$4 \leqslant k < 5$	0.01972771
5	0.05839920	$5 \leqslant k < 6$	0.07812691
6	0.14599800	$6 \leqslant k < 7$	0.22412491
7	0.25028229	$7 \leqslant k < 8$	0.47440720
8	0.28156757	$8 \leqslant k < 9$	0.75597477
9	0.18771172	$9 \leqslant k < 10$	0.94368649
10	0.05631351	$10 \leqslant k$	1.00000000
other	0		





Using pdf and cdf table (binomial $n = 10$, $p = 3/4$)

Different inequality symbols $\leqslant, >, <, \geqslant$

k	pdf $p_X(k)$		cdf $F_X(k)$
		$k < 0$	0
0	0.00000095	$0 \leqslant k < 1$	0.00000095
1	0.00002861	$1 \leqslant k < 2$	0.00002956
2	0.00038624	$2 \leqslant k < 3$	0.00041580
3	0.00308990	$3 \leqslant k < 4$	0.00350571
4	0.01622200	$4 \leqslant k < 5$	0.01972771
5	0.05839920	$5 \leqslant k < 6$	0.07812691
6	0.14599800	$6 \leqslant k < 7$	0.22412491
7	0.25028229	$7 \leqslant k < 8$	0.47440720
8	0.28156757	$8 \leqslant k < 9$	0.75597477
9	0.18771172	$9 \leqslant k < 10$	0.94368649
10	0.05631351	$10 \leqslant k$	1.00000000
other	0		

- $P(X \leqslant 2) = 0.00041580$
- $P(X > 2) = 1 - P(X \leqslant 2) = 1 - 0.00041580 = 0.99958420$
- $P(X < 2) = P(X \leqslant 2^-) = F_X(2^-) = 0.00002956$
  using infinitesimal notation from Calculus: 2^- is just below 2.
- $P(X \geqslant 2) = 1 - P(X < 2) = 1 - F_X(2^-) = 0.99997044$

Using pdf and cdf table (binomial $n = 10$, $p = 3/4$)

Probability of an interval

k	pdf $p_X(k)$		cdf $F_X(k)$
		$k < 0$	0
0	0.00000095	$0 \leqslant k < 1$	0.00000095
1	0.00002861	$1 \leqslant k < 2$	0.00002956
2	0.00038624	$2 \leqslant k < 3$	0.00041580
3	0.00308990	$3 \leqslant k < 4$	0.00350571
4	0.01622200	$4 \leqslant k < 5$	0.01972771
5	0.05839920	$5 \leqslant k < 6$	0.07812691
6	0.14599800	$6 \leqslant k < 7$	0.22412491
7	0.25028229	$7 \leqslant k < 8$	0.47440720
8	0.28156757	$8 \leqslant k < 9$	0.75597477
9	0.18771172	$9 \leqslant k < 10$	0.94368649
10	0.05631351	$10 \leqslant k$	1.00000000
other	0		

$$F_X(4) = P(X \leqslant 4) = p_X(0) + p_X(1) + p_X(2) + p_X(3) + p_X(4)$$

$$F_X(2) = P(X \leqslant 2) = p_X(0) + p_X(1) + p_X(2)$$

$$\begin{aligned} P(2 < X \leqslant 4) &= p_X(3) + p_X(4) \\ &= P(X \leqslant 4) - P(X \leqslant 2) = F_X(4) - F_X(2) \\ &= 0.01972771 - 0.00041580 = 0.01931191 \end{aligned}$$

Using pdf and cdf table (binomial $n = 10$, $p = 3/4$)

Converting other inequalities to the form $P(a < X \leqslant b)$

k	pdf $p_X(k)$		cdf $F_X(k)$
		$k < 0$	0
0	0.00000095	$0 \leqslant k < 1$	0.00000095
1	0.00002861	$1 \leqslant k < 2$	0.00002956
2	0.00038624	$2 \leqslant k < 3$	0.00041580
3	0.00308990	$3 \leqslant k < 4$	0.00350571
4	0.01622200	$4 \leqslant k < 5$	0.01972771
...	...	...	...

The formula $P(a < X \leqslant b) = F_X(b) - F_X(a)$ uses $a < X$ (not $a \leqslant X$) and $X \leqslant b$ (not $X < b$). Other formats must be converted to this.

- $$P(2 < X \leqslant 4) = P(X \leqslant 4) - P(X \leqslant 2) = F_X(4) - F_X(2) = 0.01972771 - 0.00041580 = 0.01931191$$
- $$P(2 \leqslant X \leqslant 4) = P(2^- < X \leqslant 4) = F_X(4) - F_X(2^-) = 0.01972771 - 0.00002956 = 0.01969815$$
- $$P(2 < X < 4) = P(2 < X \leqslant 4^-) = F_X(4^-) - F_X(2) = 0.00350571 - 0.00041580 = 0.00308991$$
- $$P(2 \leqslant X < 4) = P(2^- < X \leqslant 4^-) = F_X(4^-) - F_X(2^-) = 0.00350571 - 0.00002956 = 0.00347615$$

Using pdf and cdf table

Probability of an interval for integer random variables

- **Summary:** To compute the probability of an interval, convert one-sided inequalities to $P(X \leqslant b) = F_X(b)$ and two-sided inequalities to $P(a < X \leqslant b) = F_X(b) - F_X(a)$.

- We did the conversion with infinitesimals: $P(X < 2) = P(X \leqslant 2^-) = F_X(2^-) = 0.00002956$.

- **Another method:** The binomial distribution X only has integer values, so $P(X < b) = P(X \leqslant b-1)$ for any integer b. Don't use this method when non-integer values are possible.

- $P(X < 2) = P(X \leqslant 1) = F_X(1) = 0.00002956$

- $$\begin{aligned} P(2 \leqslant X \leqslant 4) &= P(1 < X \leqslant 4) = F_X(4) - F_X(1) \\ &= 0.01972771 - 0.00002956 = 0.01969815 \end{aligned}$$

- $$\begin{aligned} P(2 < X < 4) &= P(2 < X \leqslant 3) = F_X(3) - F_X(2) \\ &= 0.00350571 - 0.00041580 = 0.00308991 \end{aligned}$$

Discrete is not equivalent to integer!

- **New example, not the same as the previous example:** Suppose the range of Y is $\{0.0, 0.1, 0.2, \ldots, 9.9, 10.0\}$.
- This range is not integers, but is discrete.
- Don't convert $P(Y < a)$ into $P(Y \leqslant a - 1)$. Instead, convert it to $P(Y \leqslant b)$, where b is the largest element below a that's in the range.
- $P(Y < 2) = P(Y \leqslant 1.9)$

  $P(2 \leqslant Y \leqslant 4) = P(1.9 < Y \leqslant 4) = F_Y(4) - F_Y(1.9)$