

# 3.1–3.3 Binomial Distribution and Discrete Random Variables

Prof. Tesler

Math 186  
January 23, 2009

*Note:* Skip “The Hypergeometric Distribution” on pp. 138–146.

# Random variables

- A *random variable*  $X$  is a function assigning a real number to each outcome in a sample space.

- Flip a 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  $p_X(k) = P(X = k)$ :

$$p_X(0) = q^3 \quad p_X(1) = 3pq^2 \quad p_X(2) = 3p^2q \quad p_X(3) = p^3$$

- $p_X(k)$  is defined for *all* real numbers  $k$ .

In this case,  $p_X(k) = 0$  for  $k \neq 0, 1, 2, 3$ :

$$p_X(4) = 0 \quad p_X(2.5) = 0 \quad p_X(-3) = 0 \quad p_X(\pi) = 0 \quad \dots$$

# 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*.
- **Notation**  $p_X(k) = P(X = k)$ : Use capital letters ( $X$ ) for random variables and lowercase ( $k$ ) to stand for numeric values.
- A discrete probability density function requires  $p_X(k) \geq 0$  for all  $k$ , and that the total probability is  $\sum_k p_X(k) = 1$ .

# Binomial distribution

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

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

# Binomial distribution

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

- The range of  $X$  is  $0, 1, 2, \dots, n$ .
- $p_X(k) \geq 0$  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

| <b>Genotype</b> | <b>Probability</b> | <b>Phenotype</b> |
|-----------------|--------------------|------------------|
| $TT$            | 1/4                | tall             |
| $Tt$            | 1/2                | tall             |
| $tt$            | 1/4                | short            |

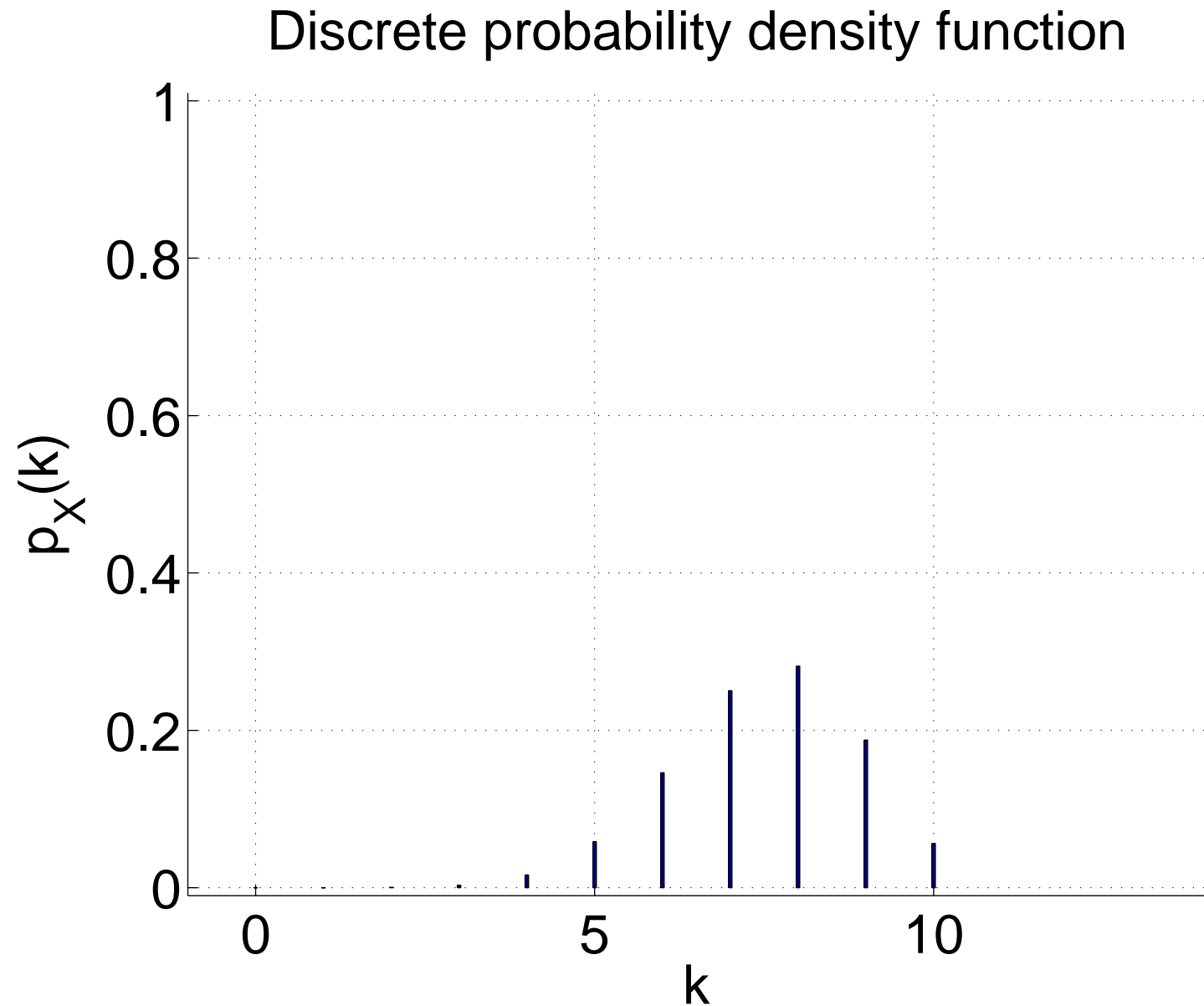
so the phenotypes have  $P(\text{tall}) = 3/4$ ,  $P(\text{short}) = 1/4$ .

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

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

# 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

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

defined over *all* real numbers  $k$ .

- For the binomial distribution,

$$F_X(k) = \begin{cases} 0 & \text{if } k < 0; \\ \sum_{i=0}^{\lfloor k \rfloor} p_X(i) = \sum_{i=0}^{\lfloor k \rfloor} \binom{n}{i} p^i (1-p)^{n-i} & \text{if } 0 \leq k \leq n; \\ 1 & \text{if } k > n. \end{cases}$$

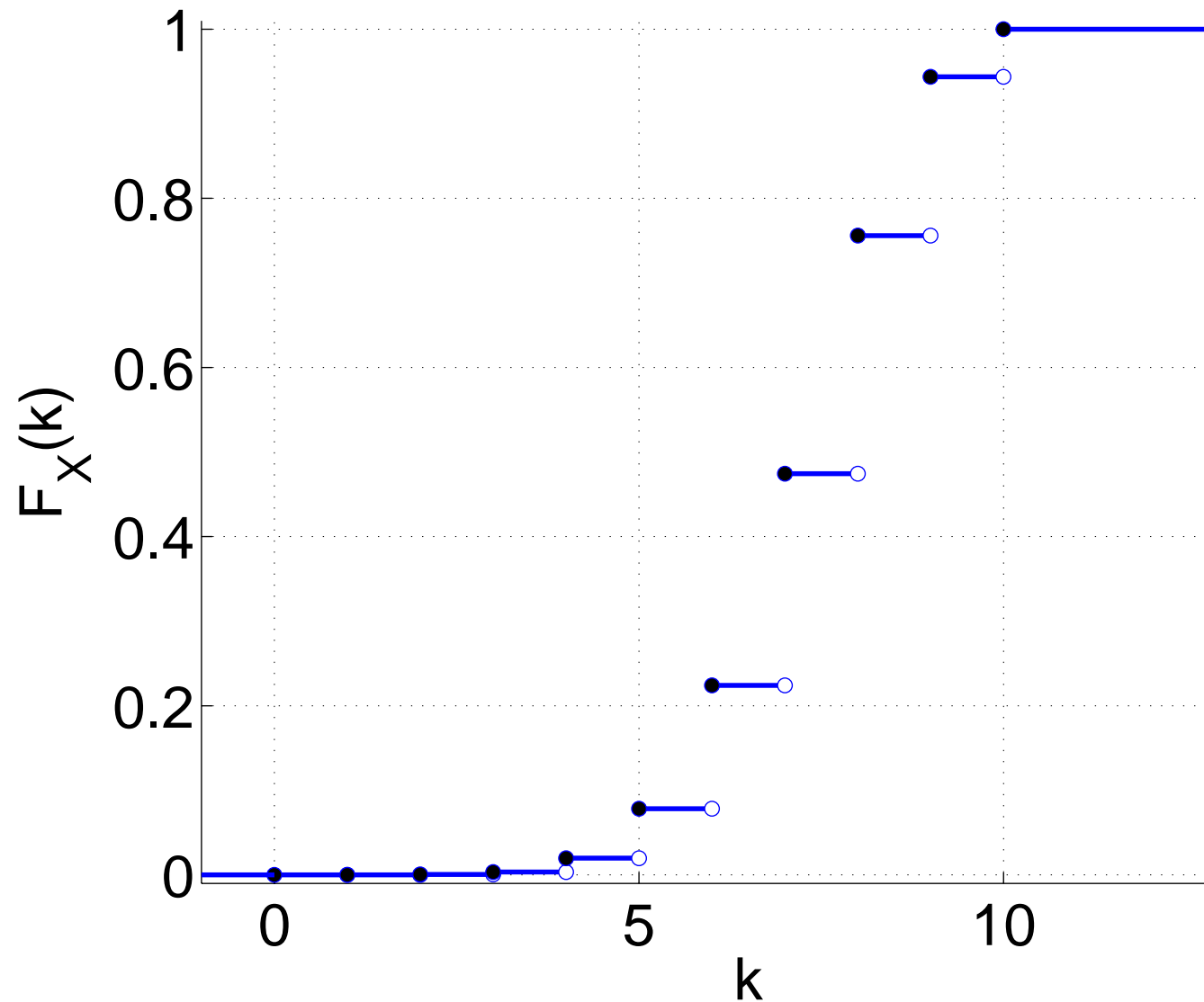
where  $\lfloor k \rfloor$  is the *floor of  $k$*  (largest integer  $\leq k$ ):

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

# CDF of binomial distribution for $n = 10, p = 3/4$

| $k$             | cdf $F_X(k)$ |
|-----------------|--------------|
| $k < 0$         | 0            |
| $0 \leq k < 1$  | 0.00000095   |
| $1 \leq k < 2$  | 0.00002956   |
| $2 \leq k < 3$  | 0.00041580   |
| $3 \leq k < 4$  | 0.00350571   |
| $4 \leq k < 5$  | 0.01972771   |
| $5 \leq k < 6$  | 0.07812691   |
| $6 \leq k < 7$  | 0.22412491   |
| $7 \leq k < 8$  | 0.47440720   |
| $8 \leq k < 9$  | 0.75597477   |
| $9 \leq k < 10$ | 0.94368649   |
| $10 \leq k$     | 1.00000000   |

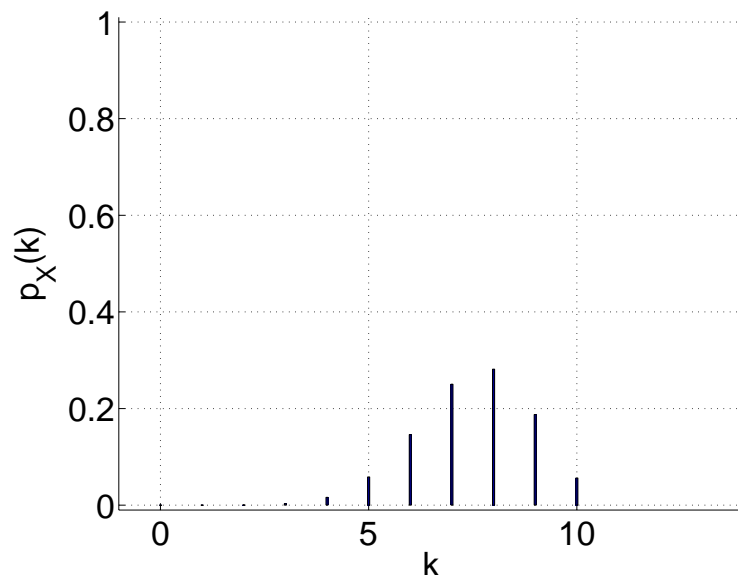
Cumulative distribution function



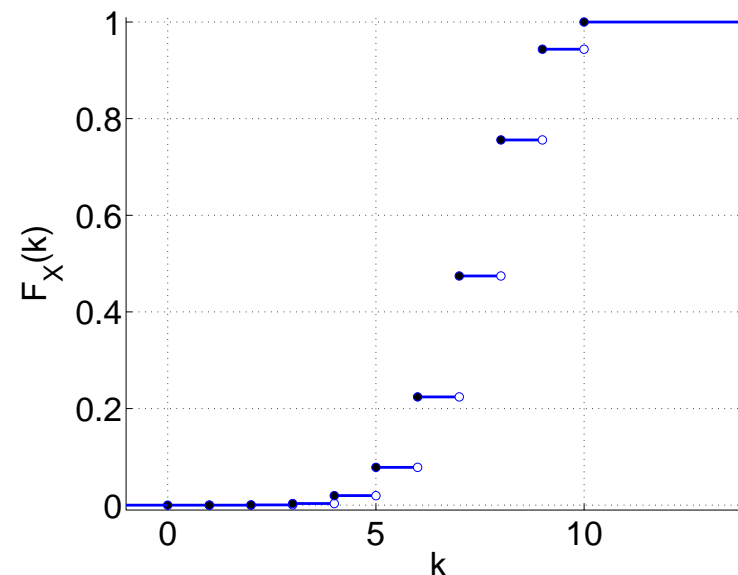
# Binomial distribution for $n = 10, p = 3/4$

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

Discrete probability density function



Cumulative distribution function



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

Plugging in arbitrary real numbers

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

- $F_X(.9999) = P(X \leq .9999) = p_X(0) = .00000095$
- $F_X(1) = P(X \leq 1) = p_X(0) + p_X(1) = .00000095 + .00002861 = .00002956$
- $F_X(1.5) = P(X \leq 1.5) = p_X(0) + p_X(1) = .00002956$
- $p_X(1.5) = 0$

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

Plugging in numbers outside of the range

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

- $F_X(-3.2) = P(X \leq -3.2) = 0$  since minimum  $X$  in range is 0.
- $F_X(12.8) = P(X \leq 12.8) = 1$  since the whole range is  $\leq 12.8$ .
- For any random variable:  $\lim_{k \rightarrow -\infty} F_X(k) = 0$        $\lim_{k \rightarrow +\infty} F_X(k) = 1$
- As  $k$  goes from  $-\infty$  to  $\infty$ , the cdf weakly increases (with jumps).

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

Different inequality symbols  $\leq, >, <, \geq$

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

- $P(X \leq 2) = 0.00041580$
- $P(X > 2) = 1 - P(X \leq 2) = 1 - 0.00041580 = 0.99958420$
- $P(X < 2) = P(X \leq 2^-) = F_X(2^-) = 0.00002956$
- $P(X \geq 2) = 1 - P(X < 2) = 1 - F_X(2^-) = 0.99997044$

# 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   | $0 \leq k < 1$  | 0.00000095   |
| 1     | 0.00002861   | $1 \leq k < 2$  | 0.00002956   |
| 2     | 0.00038624   | $2 \leq k < 3$  | 0.00041580   |
| 3     | 0.00308990   | $3 \leq k < 4$  | 0.00350571   |
| 4     | 0.01622200   | $4 \leq k < 5$  | 0.01972771   |
| 5     | 0.05839920   | $5 \leq k < 6$  | 0.07812691   |
| 6     | 0.14599800   | $6 \leq k < 7$  | 0.22412491   |
| 7     | 0.25028229   | $7 \leq k < 8$  | 0.47440720   |
| 8     | 0.28156757   | $8 \leq k < 9$  | 0.75597477   |
| 9     | 0.18771172   | $9 \leq k < 10$ | 0.94368649   |
| 10    | 0.05631351   | $10 \leq k$     | 1.00000000   |
| other | 0            |                 |              |

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

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

$$P(2 < X \leq 4) = p_X(3) + p_X(4)$$

$$= P(X \leq 4) - P(X \leq 2) = F_X(4) - F_X(2)$$

$$= 0.01972771 - 0.00041580 = 0.01931191$$

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

- $$P(2 < X \leq 4) = P(X \leq 4) - P(X \leq 2) = F_X(4) - F_X(2)$$

$$= 0.01972771 - 0.00041580 = 0.01931191$$
- $$P(2 \leq X \leq 4) = P(2^- < X \leq 4) = F_X(4) - F_X(2^-)$$

$$= 0.01972771 - 0.00002956 = 0.01969815$$
- $$P(2 < X < 4) = P(2 < X \leq 4^-) = F_X(4^-) - F_X(2)$$

$$= 0.00350571 - 0.00041580 = 0.00308991$$
- $$P(2 \leq X < 4) = P(2^- < X \leq 4^-) = F_X(4^-) - F_X(2^-)$$

$$= 0.00350571 - 0.00002956 = 0.00347615$$

# 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  $P(X \leq b) = F_X(b)$  and two-sided inequalities to  $P(a < X \leq b) = F_X(b) - F_X(a)$ .
- We did the conversion with infinitesimals:  
$$P(X < 2) = P(X \leq 2^-) = F_X(2^-) = 0.00002956.$$
- Since  $X$  only has integer values, we can also use  $P(X < b) = P(X \leq b - 1)$  (provided  $b$  is an integer).
- $$P(X < 2) = P(X \leq 1) = F_X(1) = 0.00002956$$
- $$\begin{aligned} P(2 \leq X \leq 4) &= P(1 < X \leq 4) = F_X(4) - F_X(1) \\ &= 0.01972771 - 0.00002956 = 0.01969815 \end{aligned}$$
- $$\begin{aligned} P(2 < X < 4) &= P(2 < X \leq 3) = F_X(3) - F_X(2) \\ &= 0.00350571 - 0.00041580 = 0.00308991 \end{aligned}$$